TWO-STAGE GENOME SEARCH DESIGN IN AFFECTED-SIB-PAIR METHOD

By

CHI-HSE TENG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1997

© Copyright 1997

by

CHI-HSE TENG

ACKNOWLEDGMENTS

I would like to thank the chairman of my committee, Professor Mark C.K. Yang, for introducing me to the field of genetic statistics and his wise guidance throughout my graduate study at UF. I also want to thank Professor Randy Carter, Professor Frank Martin, Professor Sue McGorrary, and Professor Jin Xiong She. Their advice, encouragement, and patience were essential. Last but not least, I want to thank Ms. Margaret Joyner for her gracious editorial help and effort. Finally, my deepest appreciation goes to my parents, brother, sister, and parents-in-law for their love and never-ending support and to my wife Hsiao-Yun and our daughter Gillian for their sweet company.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iii

LIST OF TABLES vi

ABSTRACT viii

CHAPTERS

1 INTRODUCTION 1

2 LITERATURE REVIEW 8

2.1 Map Function 8
2.2 An Introduction to Human Linkage Data for Genetic Dissection 11
2.3 Identical by State, Identical by Descent 16
2.4 Distributions of IBD 17
2.5 Risk Ratio 24
2.6 Test Statistics Based on IBD Score 25
2.7 Likelihood Ratio Test 30
2.8 Heterogeneity and Homogeneity 33
2.9 Polygenes 37
2.10 Two-stage Genome Search 38

3 TWO-STAGE GENOME SEARCH FOR SIMPLE MENDELIAN DISEASE 40

3.1 Assumptions 41
3.2 Two-Stage Procedure 42
3.3 Probability of Allocating the Correct Marker 43
3.4 Type I Error and Power of Claiming Linkage 72
3.5 Discussion 76

4 TWO-STAGE GENOME SEARCH FOR COMPLEX DISEASE 78

4.1 Genetic Model and Assumptions 78

4.2 Two-stage Genome Search 79
4.3 Probability of Allocating the Correct Marker for a Complex Disease 81
4.4 Discussion 110

5 CONCLUDING REMARKS 112

REFERENCES 114

BIOGRAPHICAL SKETCH 120

LIST OF TABLES

Table		Page
1.1	Results of Mendel's crosses experiments.	2
1.2	The results of Bateson and Bunnet's experiment.	5
2.1	Multilocus-feasibility of several map functions.	12
2.2	Some sampling issues in linkage studies.	16
2.3	The values of Pr(IBD of trait gene | IBD of marker and relationship).	20
2.4	The probability of IBD score for different relatives.	26
3.1	The joint distribution of IBD score of the locus 1 and locus 2.	57
3.2	The joint distribution of X_{ij} and $X_{i'j}$.	58
3.3	Optimal resource allocation in two-stage genome search for rare recessive gene.	64
3.4	Optimal resource allocation in two-stage genome search for rare dominant gene.	68
3.5	The 95% percentile of the unique maximum of R Binomial(n_2, 0.25).	74
3.6	The 95% percentile of the unique maximum marker group of R Binomial(n_2, 0.25).	75
4.1	Trait IBD distribution conditional on parental genotype and E vector.	83
4.2	Exclusive events for the case "G_1 is found but G_2 is not."	84
4.3	Exclusive events for the case "G_2 is found but G_1 is no."	85
4.4	Exclusive events for the case "G_1 and G_2 are found."	85
4.5	Conditional distribution of X_{1j} and X_{2j} given I_1.	87

4.6 P(both affected$|E_1,\ E_2$) and possible trait IBD given E_1 and E_2. 89

4.7 Optimal resource allocation in two-stage genome search for rare recessive gene, assumed two genes. 95

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

TWO-STAGE GENOME SEARCH DESIGN IN AFFECTED-SIB-PAIR METHOD

By

CHI-HSE TENG

December, 1997

Chairman: Mark C.-K. Yang
Major Department: Statistics

Since Penrose proposed the sib-pair method in 1935 and the affected-sib-pair (ASP) method in 1950 and 1953, these methods have been widely used in linkage studies. Combined with contemporary DNA-level genetic markers technology, ASP method can now be used for genome-wide searches for disease genes.

A two-stage search involves first a screening search to eliminate nonviable marker loci, then an intensive search to identify gene location using the remaining markers. Although this approach has been suggested in the literature, its properties have not been thoroughly investigated. The spacing between markers, the number of ASPs to be used in the experiment of each stage, and the criteria for markers to pass the first stage need to be determined. The major difficulties of the two-stage approach are (1) the joint distribution of the nonindependent statistics is difficult to handle; (2) the "random" dependence structure of the test statistics in the second stage; and (3) in most practically available sample sizes the high number of ties in test scores make asymptotic approaches inappropriate. This paper intends to provide a solution for designing an optimal design for rare autosomal recessive and dominant diseases.

Power computations using the multinomial distribution supported by simulation show that in most cases a two-stage design is usually better than a one stage design, but not always. Combining data from two stages will gain a small increase in accuracy. The optimal designs for resource allocation in each of the two stages are obtained and presented in tables.

CHAPTER 1
INTRODUCTION

Modern genetics began with the work of Gregor Mendel (1822-84) conducted between 1856 and 1863, forming the basis for his 1866 paper (Klug and Cummings, 1997). Gregor Mendel, an Austrian monk, performed a series of experiments on the garden peas. Based on the results of these experiments, he proposed a **particulate inheritance** theory which hypothesized that heritable biological characteristics were carried and controlled by individual "units." We call these units **genes** today, but Mendel called each unit a "Merkmal," the German word for "Character" (Levine and Miller, 1991).

In Mendel's garden pea experiments, he studied seven characters of garden peas. They are: round or wrinkled ripe seeds, yellow or green seed interiors, purple or white petals, inflated or pinched ripe pods, green or yellow unripe pods, axial or terminal flowers, and long or short stems (Griffiths et al, 1993). For each one of these seven characters, Mendel obtained pure lines of plants. "A **Pure line** is a population that all offspring produced by selfing or crossing within the population show the same form of the character being studied" (Griffiths et al., 1993). The **parental generation**, denoted as F_0, in Mendel's experiment are the plants of these pure lines. Mendel first studied these characters separately; he crossed two pure lines, one for each phenotype, of every characters to obtain the **first filial generation**, denoted as F_1. All the individuals of F_1 have only one phenotype of each characters. Next, Mendel self-fertilized F_1 and obtained **second filial generation**, denoted as F_2. The results are showed in Table 1.1. Mendel established two principles to explain the pattern of the

Table 1.1. Results of Mendel's crosses experiments.

Parental phenotype	F_1	F_2	F_2 Ratio
Round × wrinkled seeds	All round	5474 round; 1850 wrinkled	2.96:1
Yellow × green seeds	All yellow	6022 yellow; 2001 green	3.01:1
Purple × white petals	All purple	705 purple; 224 white	3.15:1
Inflated × pinched pods	All inflated	882 inflated; 299 pinched	2.95:1
Green × yellow pods	All green	428 green; 152 yellow	2.82:1
Axial × terminal flowers	All axial	651 axial; 207 terminal	3.14:1
Long × short stems	All long	787 long; 277 short	2.84:1

Source: Griffiths et al., 1993, p. 23.

data (Levine and Miller, 1991; Griffiths et al., 1993):

- The characteristics of an organism are determined by individual units of heredity called genes. Each adult organism has two alleles for each gene, one from each parent. These alleles are segregated (separated) from each other when reproductive cells are formed, each gametes receive one of the two alleles with equal chance. This is the **principle of segregation**, also being known as the **Mendel's First Law**.

- In an organism with contrasting alleles for the same gene, one allele may be dominant over another. This is known as the **principle of dominance**.

He assumed that each plant had two alleles for each trait his studied. This assumption was correct. "A **Allele** is one of the different forms of a gene that can exist at a single locus" (Griffiths et al., 1993, p. 783). He introduced the notation **A** for the dominant allele and **a** for the recessive allele. Since the parental generations were pure lines, their genotypes were homozygous **AA** and **aa**. Thus, the F_1 population were all heterozygous **Aa** phenotype. The F_2, offspring of the F_1, were expected to

be **AA**, **Aa**, and **aa** in the ratio 1:2:1; and dominant (**AA** and **Aa**) and recessive (**aa**) phenotype in ratio 3:1. Mendel cultivated 10 seeds from each of 100 dominant F_2 plants, if all 10 offspring from a single F_2 plant were of the dominant character then he concluded this plant was homozygous. Mendel used this experiment to demonstrate the 1:2 ratio of homozygous (**AA**) dominant and heterozygous dominant (**Aa**) F_2's. However, for a heterozygous F_2 parent there is still a $(0.75)^{10} = 0.0563$ probability would produce 10 dominant offspring, and hence been misclassified as homozygous. Therefore, the true expected ratio should be 0.3709:0.6291 instead of 1:2. Fisher suggested that Mendel's data are too close to 1:2 rather than the correct values, and thus suspected there was some manipulation, or omission of data (Weir, 1996).

Mendel also studied the two characters together; for example the shape of the seeds and the color of the seed. We denote round allele as **R** and wrinkled allele as **r**, and yellow allele as **Y** and green allele as **y**. He crossed a parental line of plants with yellow and round seeds (genotype **RRYY**) with a line with green and wrinkled seeds (genotype **rryy**). The seeds of F_1 were all round and yellow (**RrYy**) showing that these two alleles, **R** and **Y**, are dominant over **r** and **y**. Then, he crossed F_1 and got four types of seeds, 315 round/yellow seeds, 108 round/green, 101 wrinkled/yellow, and 32 wrinkled/green. The ratio was approximately 9:3:3:1 (Levine and Miller, 1991). All possible genotype combinations of F_2 are given in the following **Punnett Square** (Griffiths et al., 1993); the first row represents the genotype of gamete from father, and the first column represents the genotype of gamete from mother, and the numbers in the parenthesis are the expected frequencies if the alleles of these two characters are segregated independently.

	RY ($\frac{1}{4}$)	Ry ($\frac{1}{4}$)	ry ($\frac{1}{4}$)	rY ($\frac{1}{4}$)
RY ($\frac{1}{4}$)	RRYY ($\frac{1}{16}$)	RRYy ($\frac{1}{16}$)	RrYy ($\frac{1}{16}$)	RrYY ($\frac{1}{16}$)
Ry ($\frac{1}{4}$)	RRYy ($\frac{1}{16}$)	RRyy ($\frac{1}{16}$)	Rryy ($\frac{1}{16}$)	RrYy ($\frac{1}{16}$)
ry ($\frac{1}{4}$)	RrYy ($\frac{1}{16}$)	Rryy ($\frac{1}{16}$)	rryy ($\frac{1}{16}$)	rrYy ($\frac{1}{16}$)
rY ($\frac{1}{4}$)	RrYY ($\frac{1}{16}$)	RrYy ($\frac{1}{16}$)	rrYy ($\frac{1}{16}$)	rrYY ($\frac{1}{16}$)

Mendel established the following principle which could explain the pattern of the data.

- Each of the seven genes that Mendel investigated segregated independently (independent assortment), also being known as the **Mendel's Second Law** (Levine and Miller, 1991; McPeek, 1997).

Regardless of the controversy of his data, Mendel's theory regarding genes controlling biological characters is well accepted.

Correns (1900) observed the phenomenon of **complete linkage** in which alleles of two or more different characters appeared to be always inherited together, rather than independently as Mendel's Second Law (McPeek, 1997). This violation led to an extension of Mendel's theory; the **chromosome theory of heredity** which says that the genes are parts of specific cellular structures, the chromosomes (Griffiths et al., 1993). "In 1902, Walter Sutton and Theodor Boveri, independently published papers linking their discoveries of the behavior of chromosomes during meiosis to the Mendel's principles of segregation and independent assortment. Sutton and Boveri are credited with initiating the chromosomal theory of heredity" (Klug, 1997, p. 61). This theory provided a physical mechanism for Mendel's Laws. It were assumed that those characters that Mendel studied lay on different chromosomes, and that those which were completely linked lay on the same chromosome (McPeek, 1997).

Bateson and Punnett did an experiment on the sweet peas to study two characters: the flower color, purple (dominant) and red (recessive), and the form of pollen, long

(dominant) and round (recessive). First, they crossed purple, long (**PPLL**) with red, round (**ppll**) to create a heterozygotes (**PpLl**) type (F1 generation), then they crossed F1 generation among themselves and produced 381 plants. Result is as follows (Levine, 1991, p. 193). These results did not match Mendel's Law, nor these two

Table 1.2. The results of Bateson and Bunnet's experiment.

Phenotype	Number observed	Percent observed	Percent expected if unlinked	Percent expected if completely linked
Purple, long	284	74%	56%	75%
Purple, round	21	6%	19%	-
Red, long	21	6%	19%	-
Red, round	55	14%	6%	25%

Source: Levine, 1991.

genes were completely linked. Morgan (1911) provided an explanation for Bateson and Punnett's observation and for his own fruit fly experiment's results which is similar to Bateson and Punnett's. He suggested the exchange of genetic material had occurred between two homologous chromosomes when they paired during cell division, and this exchange was called a **crossover** (McPeek, 1997). **Recombination** is the process by which progeny derive a combination of genes different from that of either parent (DOE, 1992). Each individual inherits one chromosome from its father and the other one from its mother. When this individual reproduces two chromosomes to pass to its offspring during meiosis, it will not pass one of each pair that it inherited from its parents but a "blended" one. Meiosis produces crossovers that blend two chromosomes and cause recombinations. If there is no crossover or an even number of crossovers between two loci, then recombination does not happen to these two loci. On the other hand, if there are odd number of crossovers, then recombination happens. The probability of recombination happening between two loci is called

the **recombination fraction**. At the state of current technology level, cross-overs are not directly observable but recombination between two genes (or markers) are (Ott, 1991). An obvious consequence of the crossover process, is that the closer two genes are to each other on a chromosome, the less chance that crossovers can happen between them. Hence there is less chance that recombination can happen between them. This idea became the foundation of linkage analysis. Recombination is the chief source of variation among species (Rhodes et al., 1974).

The purpose of linkage study is to find the relative location between genes or markers and finally to present them in a linkage map. The distance between two loci on a chromosome is measured using genetic distance. We will discuss this in §2.1 "Map Function".

Lander and Schork (1994) summarized that there are four major categories of methods for genetic dissection. One is "genetic analysis of large crosses in model organisms such as the mouse and rat." The other three use human genetic data. These three methods are, pedigree analysis, allele-sharing methods, and association studies in human population. We will discuss more details in §2.2 "An Introduction to Human Linkage Data for Genetic Dissection". Affected-sib-pair method, proposed by Penrose (1953), is one of allele-sharing methods. Affected-sib-pair method combined with modern DNA-level genetic marker technology, like the Restriction Fragment Length Polymorphism markers (RFLP) (Botstein et al., 1980), can be used for genome-wide gene search.

This dissertation will discuss a two-stage genome-wide approach in affected-sib-pair method for searching for genetic disease genes. Only the designs for searching for rare recessive and dominant diseases with dichotomous phenotype (affected or not affected) is studied. Chapter 2 provides a literature review of related topics, including map function, sib-pair method, affected-sib-pair methods, identical by descent, identical by state, affected-relative-members method, heterogeneity, and polygeneity.

Chapter 3 discusses how to design an optimal two-stage search for a simple Mendelian disease gene and Chapter 4 discusses how to design an optimal two-stage search for two low penetrance, no interaction, unlinked recessive, complex disease genes. Brief concluding remarks are given in the last chapter.

CHAPTER 2
LITERATURE REVIEW

In this chapter, the following topics will be covered:

- map function (without biological interference),
- concepts of identical-by-descent (IBD) and identical-by-state (IBS),
- test statistics; based on IBD, IBS scores and likelihood ratio,
- sib-pair method and affected-sib-pair method,
- heterogeneity and homogeneity,
- polygeneity,
- two-stage approach.

2.1 Map Function

The genetic **map distance**, in units called Morgans, between two genes is defined as the expected number of crossovers occurring between two genes on a single chromosome strand (Ott, 1991). When the distance is so small that the probability of multiple crossovers is negligible, the genetic map distance is equivalent to the recombination fraction. A centimorgan between two genes is a distance that produces an approximate 0.01 probability that recombination taking place between them. When the distance is large, the probability of multiple crossovers increases. If there is zero or an even number of crossovers, then we will not observe recombination, but if there

is an odd number of crossovers we will observe recombination. Hence, the recombination fraction is not necessarily equal to the map distance when two loci are farther apart. A **map function** is used to relate the additive, but hard to measure, genetic or map distance to the non-additive, but more readily estimable, recombination fraction (Speed, 1997). Map function should preserve the additivity of map distance but not every map function does. For example, consider three loci, x_1, x_2, x_3 on a chromosome and let θ_{12} and θ_{23} be the map distance between x_1 and x_2, and x_2 and x_3 respectively, and r_{12} and r_{23} be the recombination fraction distance between x_1 and x_2, and x_2 and x_3 respectively. If we assume that x_1, x_2, and x_3 are very close such that the chance of multiple crossovers between them is negligible then the Morgan map function will not preserve additive of distance but the Haldane (1911) map function will. For Morgan map function, $r_{12} = \theta_{12}$ and $r_{23} = \theta_{23}$, the map distance between x_1 and x_3, θ_{13} is equal to $\theta_{12} + \theta_{23}$, hence recombination fraction between x_1 and x_3, r_{13} is equal to $r_{12} + r_{23}$. On the other hand, since the chance of multiple crossovers is negligible, then

$$
\begin{aligned}
& Pr(\text{recombination happen between } x_1 \text{ and } x_3) \\
= \; & Pr(\text{recombination happen between } x_1 \text{ and } x_2 \text{ but does not between } x_2 \text{ and } x_3) \\
+ \; & Pr(\text{recombination does not happen between } x_1 \text{ and } x_2 \text{ but between } x_2 \text{ and } x_3) \\
= \; & \theta_{12}(1 - \theta_{23}) + (1 - \theta_{12})\theta_{23} \\
= \; & \theta_{12} + \theta_{23} - 2\theta_{12}\theta_{23} \\
\neq \; & r_{12} + r_{23}.
\end{aligned}
$$

This is an contradiction. For Haldane map function, since the map distance is $\theta_{12}+\theta_{23}$, r_{13} is equal to $0.5(1 - \exp^{-2(\theta_{12}+\theta_{23})})$, and

$$
\begin{aligned}
& Pr(\text{recombination happen between } x_1 \text{ and } x_3) \\
= \; & 0.5(1 - \exp^{-2\theta_{12}})[1 - 0.5(1 - \exp^{-2\theta_{23}})] + 0.5(1 - \exp^{-2\theta_{23}})[1 - 0.5(1 - \exp^{-2\theta_{12}})]
\end{aligned}
$$

$$= 0.5\{(1-\exp^{-2\theta_{12}}) + (1-\exp^{-2\theta_{23}}) - (1-\exp^{-2\theta_{12}})(1-\exp^{-2\theta_{23}})\}$$

$$= 0.5(1-\exp^{-2(\theta_{12}+\theta_{23})}).$$

The additive of map distance is preserved.

Also, "most map functions are determined by three loci comparisons which may not be consistent in terms of four or more loci" (Liberman and Karlin, 1984). If a map function is consistent for any number of loci then it is called multilocus-feasible. Several counterexamples in Ott (1991, p. 127) showed that a non-multilocus-feasible might result a negative probability.

Liberman and Karlin (1984) investigated these problems. Their conclusions can be summarized as follows: there are two methods for constructing a genetic map function, the first starts with a model of the recombination process, the second uses a differential equation method. "The renewal crossover formation process model assumes that crossovers occur in succession along the chromosome, starting at a natural biological site (e.g., the centromere, some origin of replication, or a telomere) such that the lengths of the intervals between successive crossovers are positive random variables independently and identically distributed."

For the first method, if a renewal crossover formation model is specified, the only feasible map function is the Haldane map function (Haldane, 1919). The corresponding crossover formation process is a Poisson process.

Theorem 1 Consider a crossover formation model of the form of a renewal process with intercrossover distribution F(t), the distribution function of the distance between two successive crossovers, such that $\frac{d^n \mathrm{F}}{dt^n}(0) \neq 0$ for some positive integer n. Then a map function exists if and only if the intercrossover distribution agrees with that of an exponential distribution, i.e., the renewal process is a Poisson process (Liberman and Karlin, 1984).

On the issue of multilocus feasibility of map functions, they derived necessary and sufficient conditions for multilocus-feasibility.

Theorem 2 Let r be recombination fraction and x be genetic distance. Sufficient conditions: A map function $r = M(x)$ is multilocus-feasible if its derivative functions $M^{(n)}$ obey the inequalities

$$(-1)^n M^{(n)}(x) \leq 0, \qquad n = 1, 2, \ldots, \text{ for all } x \geq 0.$$

Necessary conditions: Let $r = M(x)$ be a map function that is multilocus-feasible and suppose that all derivatives $M^{(n)}(x)$ exist. Then

$$(-1)^n M^{(n)}(0) \leq 0, \qquad \forall n = 1, 2, \ldots.$$

Table 2.1 shows the multilocus feasibility of several map functions. Note that the multilocus-feasible map functions are constructed based on the cross-over formation process method, where r is the recombination fraction and x is genetic map distance. Because further extension will incorporate genetical interference, we will not discuss it here, but it can be referred to Karlin and Liberman (1994), and Speed (1997). Most map functions constructed by a differential equation method are not multilocus feasible, also will not be discussed.

2.2 An Introduction to Human Linkage Data for Genetic Dissection

The purpose of linkage study is to locate the gene which we are interested in. With contemporary DNA-level genetic marker technology (Botstein et al., 1980), one way to achieve this is by studying the linkage between genes and markers. "When two genes are inherited independently of each other, recombinants and nonrecombinants are expected in equal proportions among the offspring. For some pairs of genes, one observes a consistent deviation from the 1:1 ratio of recombinant to nonrecombinant

Table 2.1. Multilocus-feasibility of several map functions.

Source	Map Function $r = M(x)$	Multilocus Feasible
Haldane (1919)	$\frac{1}{2}(1 - \exp^{-2x})$	Yes
Ludwig (1934)	$\frac{1}{2}\sin(2x)$	No
Kosambi (1944)	$\frac{1}{2}\tanh(2x)$	No
Carter and Falconer (1951)	$x = \frac{1}{4}\mid \tan^{-1}(2r) + \tanh^{-1}(2r) \mid$	No
Sturt (1976)	$\frac{1}{2}\left[1 - (1 - (x/L))\exp^{-x(2L-1)/L}\right]$	Yes
Rao et al. (1977)	$x = \frac{1}{6}[\ p(2p-1)(1-4p)\ln(1-2r)$ $+16p(1-p)(2p-1)\tan^{-1}(2r)$ $+2p(1-p)(8p+2)\tanh^{-1}(2r)$ $+6(1-p)(1-2p)(1-4p)r]$	No
Felsenstein (1979)	$\frac{1-\exp^{2(K-2)x}}{2(1-(K-1)\exp^{2(k-2)x})}$	No

Source: Liberman and Karlin, 1984.

offspring . . . In other words, alleles of different genes appear to be genetically coupled, and this phenomenon is called **genetic linkage**." (Ott, 1991, p. 6) A **marker** is an small identifiable physical region on a chromosome and the inheritance of this region can be monitored (DOE, 1992). The closer a marker is to the gene, the smaller the chance that they will be recombined during the DNA replication process. Therefore, if a marker has a strong correlation with the phenotype of a gene, the gene should be in proximity to the marker. There are three major categories of methods using human genetic data: pedigree analysis, allele-sharing method, and association study. Risch and Merikangas (1996) summarized these methods as follows:

In pedigree analysis, we first collect pedigrees that contain affected members. Next we propose a genetic model (location of the gene, allele frequency, mode of inheritance, and so on), say M_1, to explain the pattern observed in the data, and compare the likelihood under M_1 with the likelihood under null hypothesis M_0, which assumes no gene in the region of the marker, by the likelihood ratio,

$$\frac{L(data|M_0)}{L(data|M_1)},$$

or equivalently by logarithm of the odds (LOD) (Barnard, 1949) score,

$$\log_{10} \frac{L(data|M_1)}{L(data|M_0)}.$$

Since this method requires specifying a genetic model, it is mostly used to identify simple Mendelian trait genes.

In allele-sharing methods, we first collect relative pairs or groups (most of the time, we collect affected relatives), and try to prove that the pattern observed in the data is not due to random Mendelian segregation. Most of these methods use identical-by-decent scores, which is the number of identical copies of the markers that relatives share. The major advantage is that no specific genetic model is required to

do the test, so they are considered as nonparametric methods. These methods are mostly used for detecting non-Mendelian complex trait genes.

Association study is a case-control study. Instead of using familial data, rather it collects unrelated affected and unaffected individuals, and then compares the allele frequencies of the candidate genes or markers. If the disease gene(s) is/are one of those candidate gene or very strongly associated with some marker alleles, then the affected group will have higher frequencies of those alleles than the control group. This method is mostly used after the possible region of the genes has been narrowed down, because this method involves examining genes and/or markers on a very fine scale, i.e., examining many markers in a small area of chromosome. "Association studies seem to be of greater power than linkage studies. But of course, the limitation of association studies is that the actual gene or genes involved in the disease must be tentatively identified before the test can be performed. . . . Thus, the primary limitation of genome-wide association tests is not a statistical one but a technological one" (Risch and Merikangas, 1996, p. 1517).

Allele-sharing methods include the affected-sib-pair (ASP) method and its variants. Penrose was considered the first person to propose the sib-pair method (Conneally and Rivas, 1980; Shah and Green, 1994). In 1935, Penrose proposed a simple χ^2 test, to detect linkage between two characters. In 1938, he extended this method to "graded" characters, which means the characters could have intermediate value. In 1950, Penrose proposed the concept of affected-sib-pair method as a means for analyzing red hair phenotype and ABO blood group data. He wrote: "Accuracy is preserved and uninformative dead wood excluded if a set of sibships is selected by the presence of one of the test characters . . ." A general form of the sib-pair method was proposed by Penrose in 1953. Now, these methods have been extended to examine identical-by-descent scores (elaborated in the following sections), identical-by-state

scores of markers, and the risk ratio between two relatives. The idea of affected-sib-pair method is only allow a pair of sibs both who are affected to be included in the study. If the disease is caused by a gene then both affected sibs tend to receive the same gene. Hence, if a marker is close to the gene, then the marker will tend to segregate with the gene during meiosis. We will cover how to detect the linkage in the later sections.

The major advantages of affected-sib-pair method over pedigree analysis were summarized by Holmans and Craddock in 1997:

- Affected-sib-pair method does not require specification of a genetic model, which is important for complex disease where the mode of inheritance is unclear.
- It is generally easier to collect affected sibling data than to collect a large, multigeneration pedigree with multiple affected members.
- Affected sib-pairs are more likely to be informative for linkage than large pedigree under oligogentic epistatic models, which is plausible for a number of complex traits.

The disadvantage of affected-sib-pair method is that it is considered less powerful than traditional pedigree analysis when the genetic model can be specified. If the genetic model cannot be specified in the affected-sib-pair methods, the recombination fraction cannot be estimated.

Suarez, Rich, and Reich (1978), and Blackwelder and Elston (1985) suggested that sampling only affected sib-pairs is more powerful then sampling affected-unaffected sibs under a fixed sample size constraint. Some sampling issues in linkage studies as summarized by Gershon et al. (1994) appear in Table 2.2.

Table 2.2. Some sampling issues in linkage studies.

Issue	Pro	Con
Very large pedigree	Statistical power high under homogeneity assumption.	Extension may actually introduce heterogeneity. Very hard to find.
Medium-size pedigree series	Less likely to have heterogeneity within pedigrees. More likely to get generalizable results.	Heterogeneity between pedigrees.
Nuclear families		Lower power: very many needed if heterogeneity present.
Affected sib-pairs	Model-free	Even lower statistical power.

Source: Gershon et al., 1994.

2.3 Identical by State, Identical by Descent

IBS stands for "identical by state." It means two alleles are the same regardless of whether the alleles are copies of the same ancestral alleles or not. IBD stands for "identical by descent." Two alleles are said to be IBD if those two alleles are copies of the same ancestral alleles. Since humans are diploid, that is, having two haploid sets of chromosomes, two sibs can share 2, 1, or 0 marker alleles at a locus. Consequently, their IBD scores would be equal to 2, 1, or 0, respectively. We will discuss the distribution of IBD score in the next section. We only consider two alleles are identical-by-decent if they are copies of the same ancestral allele in the pedigree under study, i.e., if the alleles are traceable.

With the affected-sib-pair method, since the sib-pairs were selected for study on the condition that both sibs are sick with the same disease (biological character), if the disease is caused by a gene, then a marker adjacent to the disease gene has a higher probability to segregate with the gene during meiosis. Hence, this marker tends to have a higher IBD score. This method can serve as a tool to locate the trait genes.

2.4 Distributions of IBD

Under ideal conditions, in using the affected-sib-pair method, we can take IBD scores to make inferences about the recombination fraction between the trait gene(s) and the marker(s). Hence it is important to know the distribution of the IBD scores. According to Mendel's First Law, each of a parent's haplotypes have an equal chance to be passed to offspring. Therefore, if sibs were not selected conditionally on any other biological characters, then, the probability for a pair of full sibs to have a IBD score equal to 2 is 0.25, to 1 is 0.5, and to 0 is 0.25. If sibs are selected conditional on a biological character and if the marker(s) we examine is unlinked to the trait gene(s) that responsible for the biological character, then, regardless of the inheritance mode, penetrance, and allele frequency of the trait gene the probability mass function is also 0.25, 0.5, and 0.25 for IBD equal to 2, 1, and 0, respectively. Therefore, if the observed IBD score deviates from this mass function, it is an evidence that the observed IBD scores are not from a random model. Nevertheless, under specified condition the distribution of the IBD score is still worth investigating, especially to develop a more powerful test statistics or to do power analysis.

Li and Sacks (1954) developed a method using stochastic matrices to derive the joint distribution of the genotype of two relatives. They used an example to demonstrate their method: a gene with two alleles, say A and a, with population frequencies

p and $q = 1 - p$. First, obtain a conditional probability matrix $P=\{p_{ij}\}$;,

$$\begin{array}{c} \\ \underline{\text{relative 1}} \end{array} \begin{array}{c} \underline{\text{relative 2}} \\ \begin{array}{c} AA \\ Aa \\ aa \end{array} \end{array} \begin{array}{c} \begin{array}{ccc} AA & Aa & aa \end{array} \\ \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} \end{array}$$

where p_{11} is the conditional probability conditional on relative 1 having AA, that relative 2 has AA; p_{12} that relative 2 has Aa; and p_{13} that relative 2 has aa, and so on. Once such a matrix is obtained, the absolute frequencies of all different combinations of genotypes of relatives pairs can be easily obtained by multiplying corresponding row by genotype frequencies of relative 1. In their example, that would be multiplying the first row, (p_{11}, p_{12}, p_{13}) by p^2, the second row by $2pq$, and the third row by q^2. Li and Sacks' method use three basic transition probabilities matrices, I, T, and O, to construct P. The matrices, I, T, and O is defined same as P except conditional on relatives pair having both, one, or no genes identical by descent, respectively. One can see that, in their example,

$$I = \{i_{lj}\} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, T = \{t_{lj}\} = \begin{bmatrix} p & q & 0 \\ \frac{1}{2}p & \frac{1}{2} & 0 \\ 0 & p & q \end{bmatrix}, O = \{o_{lj}\} = \begin{bmatrix} p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \\ p^2 & 2pq & q^2 \end{bmatrix},$$

where $l = 1, 2, 3$ and $j = 1, 2, 3$. For the matrix I; since given that relative 1 and 2 share two genes identical by descent, then the probability of the genotype of relative 2 is AA, given the genotype of relative 1 is AA, is equal to 1, so $i_{11} = 1$. For the same reason, the diagonal elements of I, i_{ll} are equal to one, and off diagonal elements are equal to zero. Now consider the matrix T, since given that both relative share only one gene IBD, if relative 1 has genotype AA, and relative 2 also has genotype AA, relative 2 must receive one of his A alleles somewhere else other than the same

ancestor from whom relative 1 received his A allele. Since population frequency of A is p, thus t_{11} of the matrix T is equal to p. The other elements of T and O can be obtained by the same reasoning.

They showed several examples of how to use I, T, and O to construct P. The one for the parent-offspring relationship, the transition matrix $P = T$, and the grandparent-grandchild pair has transition matrix

$$P = T^2 = \frac{1}{2}T + \frac{1}{2}O,$$

and for the general parent-offspring type relatives,

$$P = T^{n+1} = \frac{1}{2}^n T + (1 - (\frac{1}{2})^n)O,$$

where $n + 1$ is the number of generations between the two relatives. For the full-sib, the matrix is

$$P = S = \frac{1}{4}I + \frac{1}{2}T + \frac{1}{4}O.$$

Campbell and Elston (1971) extended Li and Sacks' method to derive the transition probability matrices for different modes of inheritance and multiple loci. Haseman and Elston (1972) constructed the joint probability matrix for sibs. Risch (1990b) and Bishop and Williamson (1990) summarized these works and gave a table of trait IBD distribution conditional on marker IBD and relationship for sibs, grandparent-grandchild, uncle-nephew, half-sibs, and first cousins. It is adapted here in Table 2.3.

Holmans (1993) showed that for an affected sib-pair, the possible values of probability mass function $p = (p_0, p_1, 1 - p_0 - p_1)$ of marker IBD score 0, 1, and 2, is restricted in what they called "the possible-triangle", which is the intersection of $p_0 \geq 0$, $p_1 \leq \frac{1}{2}$, and $p_1 \geq 2p_0$, for all genetic models. Their proof is included.

Let p_j and z_i be the probabilities of a sib-pair sharing ibd at marker locus and trait locus respectively, θ be the recombination fraction between the marker and the

Table 2.3. The values of Pr(IBD of trait gene | IBD of marker and relationship).

IBD of marker	IBD of Trait		
	2	1	0
sibs			
2	Ψ^2	$2\Psi(1-\Psi)$	$(1-\Psi)^2$
1	$\Psi(1-\Psi)$	$(1-2\Psi+2\Psi^2)$	$\Psi(1-\Psi)$
0	$(1-\Psi)^2$	$2\Psi(1-\Psi)$	Ψ^2
grandparent-grandchild			
1		1- θ	θ
0		θ	1-θ
uncle-nephew			
1		$\Psi(1-\theta)+\frac{1}{2}\theta$	1- $\frac{1}{2}\theta$ - $\Psi(1-\theta)$
0		1- $\frac{1}{2}\theta$ -$\Psi(1-\theta)$	$\Psi(1-\theta)+\frac{1}{2}\theta$
half-sibs			
1		Ψ	1-Ψ
0		1- Ψ	Ψ
first cousins			
1		$\Psi(1-\theta)^2+\frac{1}{2}\theta^2$	1- $\frac{1}{2}\theta^2$ - $\Psi(1-\theta)^2$
0		$\frac{1}{3}\{1-\frac{1}{2}\theta^2-\Psi(1-\theta)^2\}$	$\frac{1}{3}\{2+\Psi(1-\theta)^2+\frac{1}{2}\theta^2\}$

Where θ is the recombination fraction between the marker and trait gene,

$$\Psi = \theta^2 + (1-\theta)^2.$$

trait gene, and $\Psi = \theta^2 + (1-\theta)^2$. Then,

$$\begin{aligned} p_0 &= \Psi^2 z_0 + \Psi(1-\Psi)z_1 + (1-\Psi)^2 z_2 \\ p_1 &= 2\Psi(1-\Psi)z_0 + [\Psi^2 + (1-\Psi)^2]z_1 + 2\Psi(1-\Psi)z_2 \\ p_2 &= (1-\Psi)^2 z_0 + \Psi(1-\Psi)z_1 + \Psi^2 z_2. \end{aligned}$$

From the second equation,

$$p_1 = 2\Psi(1-\Psi)(z_0 + z_1 + z_2) + [\Psi^2 + (1-\Psi)^2 + 2\Psi(1-\Psi)]z_1.$$

Since $\frac{1}{2} \leq \Psi \leq 1$, and it can be shown that $z_1 \leq \frac{1}{2}$,

$$p_1 \leq 2\Psi(1-\Psi) + (2\Psi - 1)^2/2 = \frac{1}{2}.$$

From the first and second equations

$$\begin{aligned} & p_1 - 2p_0 \\ = \; & (2\Psi - 4\Psi^2)z_0 + (2\Psi - 1)^2 z_1 + \{(1-\Psi)[2\Psi - 2(1-\Psi)]\}z_2 \\ = \; & (2\Psi - 1)^2 z_1 + 2(1-\Psi)(2\Psi - 1)(1 - z_1 - z_0) - 2\Psi(2\Psi - 1)z_0 \\ \geq \; & (2\Psi - 1)^2 z_1 + 2(1-\Psi)(2\Psi - 1) - 3(1-\Psi)(2\Psi - 1)z_1 - \Psi(2\Psi - 1)z_1 \\ = \; & 2(1-\Psi)(2\Psi - 1)(1 - 2z_1) \\ \geq \; & 0. \end{aligned}$$

This proves the "possible triangle" restriction. He also proposed a procedure for finding the maximum likelihood estimator under the above restriction. That procedure is as follows:

1. Obtain $\widehat{z} = (\widehat{z_0}, \widehat{z_1}, 1 - \widehat{z_0} - \widehat{z_1})$ from the unrestricted method which is sample frequency.

2. If $\widehat{z_1} \geq \frac{1}{2}$, remaximize $\widehat{z}$ subject to the constraint $\widehat{z_1} = \frac{1}{2}$. If the resultant $\widehat{z_0} \geq \frac{1}{4}$, then reset $\widehat{z}$ to null which is (0.25, 0.5, 0.25).

3. If $2\widehat{z_0} > \widehat{z_1}$, remaximize $\widehat{z}$ subject to the constraint $z_1 = 2z_0$. If the resultant $\widehat{z_0} \geq \frac{1}{4}$, then reset $\widehat{z}$ to null.

4. If $\widehat{z}$ is in the triangle, leave it as it is.

This concludes the procedure.

Suarez, Rice, and Reich (1978) derived a generalized sib-pair IBD score distribution in which the IBD score distribution is conditioned on the number of affected sibs for a trait with two alleles. This conditional distribution showed that "the distribution of IBD scores depends only on the additive and dominance variances and the population prevalence of the disorder" (p. 94).

One problem with using the IBD scores is, in order to establish IBD scores, marker alleles must be highly polymorphic. "Often, identity by descent cannot unequivocally be established" (Ott, 1991, p. 79). To overcome this problem with the IBD score, Lange (1986a) proposed an affected-sib-sets method based on IBS (identical by state) and at same time he extended the affected-sib-pair method to the affected-sib-set (ASS) method (Lange, 1986b). He proposed a test statistic for the affected sibs in a single nuclear family,

$$Z = \sum_{i<j} X_{ij} \tag{2.1}$$

$$X_{ij} = \begin{cases} 1 & i \text{ and } j \text{ concordant (i.e. IBS=2)}, \\ \frac{1}{2} & i \text{ and } j \text{ half-concordant (i.e. IBS=1)}, \\ 0 & i \text{ and } j \text{ discordant (i.e. IBS=0)}, \end{cases} \tag{2.2}$$

where i and j are indexes of sib-set members. Lange noted that since they "will ultimately apply the Central Limit Theorem, it suffices to derive the mean and variance of Z for a particular affected sib set." This ASS method uses population allele frequency to calculate the probabilities of all the possible parent mating types, then, conditional on the parental mating type, to calculate the probability of IBS of the

sib-pair. Lange gave the mean and variance of Z "in the limiting case of an infinite number of alleles, each of infinitesimal frequency," (this author believes that means very highly polymorphic) as follows,

$$E(Z) = s(s-1)/4, \tag{2.3}$$

$$Var(Z) = s(s-1)/16, \tag{2.4}$$

where s is the size of the sib-set. He then presented a way to combine Z statistics from different sib-sets,

$$T = \frac{\sum_s \sum_r w_{rs}(Z_{rs} - E(Z_{rs}))}{(\sum_s \sum_r w_{rs}^2 Var(Z_{rs}))^{\frac{1}{2}}}. \tag{2.5}$$

where Z_{rs} denote the Z statistic for the rth affected sib set of size s, and w_{rs} is a weighting. To assign a value to w_{rs}, he claimed if the weights w_{rs} depend only on s and not on r, and if the number of sib sets is large, then T should follow a standard normal distribution approximately. Based on Hodge's (1984) results, Lange recommended using the weights,

$$\frac{(s-1)^{\frac{1}{2}}}{Var(Z_{rs})^{\frac{1}{2}}}$$

In 1988, Weeks and Lange (1988) generalized the ASS method to pedigrees, calling the refined method the affected-pedigree-member (APM) method. In this method only those pedigree members who are both affected and typed at the marker locus enter the definition of test statistic. They modify the above test statistic T in two ways. First, modified the Z statistic to be

$$Z_{ij} = \frac{1}{4}\delta(G_{ix}, G_{jx}) + \frac{1}{4}\delta(G_{ix}, G_{jy}) + \frac{1}{4}\delta(G_{iy}, G_{jx}) + \frac{1}{4}\delta(G_{iy}, G_{jy}),$$

where i and j are the index for affected members in a pedigree, G_{ix} and G_{jx} are the maternal marker allele for member i and j respectively, G_{iy} and G_{jy} are the paternal

marker allele for member i and j respectively, and

$$\delta(G, G') = \begin{cases} 1 & G \text{ and } G' \text{ match in state} \\ 0 & G \text{ and } G' \text{ do not match in state.} \end{cases}$$

They claim this modification "permits computation of the theoretical mean and variance of the new test statistic for each pedigree by taking advantage of the theory of multiple-person ibd relation." Second, a new weighting factor

$$w_m = \frac{\sqrt{r_m - 1}}{\sqrt{Var(Z_m)}},$$

where r_m is the number of affected and typed individuals in the mth pedigree, and Z_m is the Z statistic of the mth pedigree. Later, Week and Lange (1992) proposed a multilocus extension of this APM method.

Bishop and Williamson (1990) studied the power of IBS methods on affected relative pairs and showed that several factors have a major influence on the power. These factors are, the relationship of the affected pairs, the polymorphism of the marker, the recombination distance between a trait locus and the closest marker, and the mode of inheritance of the trait.

2.5 Risk Ratio

In 1990, Risch published a series of papers addressing risk ratio methods. In the first of this series (1990a), a ratio of risk, λ_R, is defined as the ratio of the probability of being affected given a type R relatives is affected, and the population prevalence. Let the relationship subscripts as follows: M=monozygotic twin; S=sibling (or dizygotic twin); O= parent (or offspring); 2= second-degree relatives; 3=third-degree relative. For a single-locus model, two **recurrence-risk** pattern were derived;

$$\lambda_O - 1 = 2(\lambda_2 - 1) = 4(\lambda_3 - 1) \tag{2.6}$$

and

$$\lambda_M = 4\lambda_S - 2\lambda_O - 1; \tag{2.7}$$

Similar results were obtained for multiplicative, additive, and genetic heterogeneity two-locus models and a general multilocus model. Important conclusions are: "for a single-locus model, $\lambda_R - 1$ decreases by a factor of two with each degree of relationship. . . . For a multiplicative (epistasis) model, $\lambda_R - 1$ decreases more rapidly than by a factor of two with each degree of relationship. Examination of λ_R values for various classes of relatives can potentially suggest the presence of multiple loci and epistasis".

In the second paper (1990b), Risch derived the probabilities of IBD scores for a completely polymorphic marker for several relatives under the assumption that the recurrence-risk pattern (2.6) holds. According to his first paper of this series, this assumption might be violated if multiple contributing loci present. The IBD distributions were summarized in Table 2.4. He concluded that, "for diseases with large λ values and for small θ value, distant relatives offer greater power. For large θ values, grandparent-grandchild pairs are best; for small λ values, sibs are best." The third paper (1990c) took into account the effect of marker polymorphism on power. There were some errors in the third paper and Risch wrote a response to address and corrected them (Risch 1992). The feasibility of his method is dependent on whether a suitable control group is included in the study or an appropriate estimate of general population risk can be obtained.

2.6 Test Statistics Based on IBD Score

There are several test statistics based on IBD score of the sib-pair method to detect linkage. One is the **Two-allele (proportion) test statistic**, which is the number

Table 2.4. The probability of IBD score for different relatives.

Marker IBD	Probability
Affected sib pairs	
2	$\frac{1}{4} + \frac{1/4}{\lambda_S}(2\Psi - 1)[(\lambda_S - 1) + 2\Psi(\lambda_S - \lambda_O)]$
1	$\frac{1}{2} - \frac{1}{2}(2\Psi - 1)^2 \frac{1}{\lambda_S}(\lambda_S - \lambda_O)$
0	$\frac{1}{4} - \frac{1/4}{\lambda_S}(2\Psi - 1)[(\lambda_S - 1) + 2(1 - \Psi)(\lambda_S - \lambda_O)]$
Grandparent-grandchild pairs	
1	$\frac{1}{2} + \frac{1}{2}(1 - 2\theta)\frac{1}{(\lambda_O+1)}(\lambda_O - 1)$
0	$\frac{1}{2} - \frac{1}{2}(1 - 2\theta)\frac{1}{(\lambda_O+1)}(\lambda_O - 1)$
Uncle(Aunt)-Nephew(Niece)	
1	$\frac{1}{2} + \frac{1}{2}(1 - \theta)(1 - 2\theta)^2\frac{1}{(\lambda_O+1)}(\lambda_O - 1)$
0	$\frac{1}{2} - \frac{1}{2}(1 - \theta)(1 - 2\theta)^2\frac{1}{(\lambda_O+1)}(\lambda_O - 1)$
Half-sib Pairs	
1	$\frac{1}{2} + \frac{1}{2}(2\Psi - 1)\frac{1}{(\lambda_O+1)}(\lambda_O - 1)$
0	$\frac{1}{2} - \frac{1}{2}(2\Psi - 1)\frac{1}{(\lambda_O+1)}(\lambda_O - 1)$
First-cousin pairs	
1	$\frac{1}{4} + [(1 - \theta)^4 + \theta^2(1 - \theta)^2 + \frac{1}{2}\theta^2 - \frac{1}{4}]\frac{1}{(\lambda_O+3)}(\lambda_O - 1)$
0	$\frac{3}{4} - [(1 - \theta)^4 + \theta^2(1 - \theta)^2 + \frac{1}{2}\theta^2 - \frac{1}{4}]\frac{1}{(\lambda_O+3)}(\lambda_O - 1)$

Where θ is the recombination fraction between gene and marker, $\Psi = \theta^2 + (1 - \theta)^2$.

(or proportion) of sib-pairs that share two alleles, proposed by Day and Simons (1976) and Suarez, Rice, and Reich (1978). The second, the **Mean test statistic**, is the sum (or sample mean) of IBD scores, proposed by Green and Woodrow (1977). The third is to use the Pearson chi-square goodness-of-fit test (Pearson, 1900) on the distribution of IBD scores. All these tests assume that the IBD score frequencies are 0.25, 0.5 and 0.25, under unlinked conditions, and using the deviation of observed IBD score from these frequencies to detect the linkage.

Several papers discuss the merit of these test statistics. For the Two-alleles test statistics; the one proposed Day and Simons (1976) was based on the supposition that "the probability of sharing both haplotypes deviates more from the probability under the null hypothesis than does the probability of having one, or at least one, haplotype in common. One would test for the existence of the DS (disease susceptibility) gene by comparing the observed frequency of having both haplotypes in common." Suarez, Rice, and Reich (1978) suggested pooling categories IBD=0 and IBD=1, based on their generalized sib-pair IBD distribution that we mentioned in §2.4 "Distributions of IBD". They demonstrated that the affected-sib-pair that shared two alleles had the most information, noting that "with minimal loss of information, we can pool categories Pr{IBD=0,1}."

Green and Woodrow (1977) used the total number of "repeats," i.e., the total number of IBD scores, as the test criterion. They also suggested using only affected sib-pairs because the distribution of IBD scores is symmetric under null hypothesis, so the sum of "repeats" can be approximated by normal distribution quicker than other sampling scheme.

Blackwelder and Elston (1985) examined the Two-allele test statistic, the Mean test statistics, and Goodness-of-fit test statistic, by calculating the exact probabilities. They concluded that the Mean test statistic is generally more powerful than either of the other two. Schaid and Nick (1990) proposed a test statistic using a linear

combination of the number of marker alleles IBD as a test statistic. If the alternative IBD distribution can be specified, then this test statistic can be optimized. They also suggested a t_{max} test statistic which was the maximum of the Mean test statistic and the Two-alleles test statistic and showed that t_{max} was a more robust test statistic in the sense that, if the mode of inheritance (recessive, dominant, and so on) is misspecified, loss of power may not be too great.

Faraway (1993) proposed a modified chi-square test. This test is a special case of a general method, χ^2 test with restricted alternatives, given by Lehmann (1986, p. 480). Faraway showed it was more powerful than the Two-alleles, the Mean, Pearson chi-square, or Schaid and Nick's t_{max} tests for the finite sample and claimed it was the asymptotically uniformly most powerful invariant test. More details are given in the following: For testing hypothesis

$$H_0 : p_0 = p_2 = 0.25, p_1 = 0.5 \quad vs \quad H_1 : p_0 + p_1 + p_2 = 1.$$

Let the chi-square goodness of fit test statistics be

$$Y = 4n(\widehat{p_0} - 0.25)^2 + 2n(\widehat{p_1} - 0.5)^2 + 4n(\widehat{p_2} - 0.25)^2,$$

where $\widehat{p_i}$ is the sample frequency of IBD=i, n is the total sample size. Faraway showed that the possible value of IBD score distribution, p_i $i = 1, 2, 3$, is restricted in region F that is the intersection of $p_1 + p_2 \leq 1$, $p_1 \leq 1/2$, and $3p_1/2 + p_2 \geq 1$, and claimed this region is equivalent to the "possible-triangle" in Holmans (1993). Then, showed under this restriction, the $(\widehat{p_0}\prime, \widehat{p_1}\prime, \widehat{p_2}\prime)$ that maximizes statistic Y will actually turn Y into one of Two-allele test statistic, Mean test statistic, Y, and 0. The results are as follows.

Region	Test statistic
F	Y
$2\widehat{p}_2 + \widehat{p}_1 \geq 1, \widehat{p}_1 \geq 1/2$	Mean
$3\widehat{p}_1/2 + \widehat{p}_2 < 1, \widehat{p}_2 \geq 1/4$	Two-allele
$2\widehat{p}_2 + \widehat{p}_1 < 1, \widehat{p}_2 < 1/4$	0

All these discussions regarding the merit of different test statistics were based on a model where the trait locus has only two alleles. Since the distribution of IBD score was dependents on the penetrance of the genotype, as well as the frequency of the genotype, therefore, whether these results can be extend to a model that assumes three or more alleles for a disease susceptibility gene needs further investigation.

Based on the same two-allele model, and Suarez, Rice, and Reich's (1978) generalized sib-pair IBD score distribution, Knapp, Seuchter, and Baur (1994a) showed that if $f_1^2 = f_0 f_2$, the Mean test was the uniformly most powerful test in θ (recombination fraction) regardless of the allele frequency of t, where f_0 is the penetrance of the genotype tt (t is the susceptibility allele and T is the normal allele), f_1 is for the genotype Tt, and f_2 is for the genotype TT. It is clear that recessive diseases ($f_1 = f_2 = 0$) satisfied this condition, with either complete or incomplete penetrance ($f_0 = 1$ or $f_0 \leq 1$). The authors also proved the Mean test is the locally most powerful test (there is no test with a larger power for any alternative within a neighborhood of H_0 according to Rao, 1973), irrespective of the mode of inheritance. He stated, "Because uniform optimality implies local optimality, no other test than the mean test can be uniformly most powerful." In another paper (1994b) Knapp, Seuchter, and Baur presented the equivalence of the mean test and the parametric maximum LOD score analysis with an assumed recessive mode of inheritance.

2.7 Likelihood Ratio Test

The generalized likelihood ratio test has been commomly used in genetics for detecting linkage and heterogeneity. The LOD, an acronym for "logarithm of the odds ratio" Barnard (1949), is a special version of the likelihood ratio. However, the classic research in likelihood ratio test cannot be directly applied. Let's review this theorem first.

In Serfling's book "Approximation Theorems of Mathematical Statistics", (1980, p. 144), he states the following:

> **Regularity Conditions on** F. Consider Θ to be an **open** interval (not necessarily finite) in R. We assume:
>
> ($R1$) For each $\theta \in \Theta$, the derivatives . . . ,

where F is the family of distribution function, and Θ is the parameter space of F. Assuming regularity conditions, it can be shown that the maximum likelihood estimator of θ has an asymptotic normal distribution (Serfling, 1980, p. 145). With this property and **Lemma A** (Serfling, p. 153), the likelihood ratio test theorem (Serfling, p. 155) can be proven as:

Consider testing H_0: $\theta = \theta_0$, where θ_0 is a k-dimensional vector in Θ. Let

$$\Lambda_n = \frac{L(\theta_0)}{\sup_{\theta \in \Theta} L(\theta)},$$

where n is sample size. Then, under H_0, the statistics -2ln Λ_n converge in distribution to χ^2_k (Serfling, 1980).

The purpose of regularity conditions was to make sure of the existence of the Taylor expansion of the distribution functions (Serfling, p. 145). The parameter space was assumed to be an open interval in R, in fact, if θ_0 is an interior point of the parameter space Θ, the theorem holds regardless of whether the parameter space is an open interval or not.

Let

$$\text{LOD} = \log_{10} \frac{max_{\,0 \le \theta \le 0.5}\, L(\theta)}{L(\theta = 0.5)}, \tag{2.8}$$

If the genetic model is known or is assumed to be known, some people use LOD, to to test

$$H_0\text{: } \theta = 0.5 \text{ vs } H_1\text{: } \theta < 0.5,$$

where θ is the recombination fraction between a marker and a gene, and claimed $2\ln(10) \times$ LOD $\sim \chi_1^2$ (Chotai, 1984; Elston, 1994; Shute, 1988). However, this is incorrect because $\theta_0 = 0.5$ is on the boundary of the parameter space [0, 0.5]. Therefore, the maximum likelihood estimator does not have an asymptotic normal distribution. Similar errors happen when testing heterogeneity, where the parameter space is [0,1], and, under null hypothesis, monogeneity, i.e., $\alpha = 1$ is on the boundary of parameter space (Ott, 1983), where α is the proportion of families belonging to the linked group.

Ott (1991) attempted to correct this error, arguing that LOD is a "one-sided" test. Correct asymptotic distribution of maximum likelihood ratio under the condition that null parameter value is on the boundary of the parameter space was given by Self and Liang (1987). The asymptotic distribution of Eq. (2.8) is a 50:50 mixture distribution of a χ_1^2 and a mass equal to 1 at 0. To make inferences, a critical value of LOD score needs to be decided. For simple Mendelian disease, the conventional criterion for claiming linkage is LOD ≥ 3 (Morton, 1955, 1956). The probability of obtaining a LOD ≥ 3 under the null hypothese is about 0.0001. One reason for such a low significance level is that the prior probability of that gene being within a certain distance from the marker is small; and another reason is that we would rather have no linkage map than have a wrong linkage map (Morton, 1955, 1956). Ott (1994) gave a Baysian argument, asking us to consider two hypotheses, H_0: free recombination or absence of linkage ($\theta = 0.5$), and H_1: $\theta < 0.5$. The "prior" probability that a gene and a marker are within a measurable distance (40 CM) is small, and based on

Elston and Lange's result in 1975, he said the probability of a marker and a gene being within 40 cM was equal to 0.02. The posterior odds for linkage is,

$$\frac{P(H_1|\ data)}{P(H_0|\ data)} = \left(\frac{P(data|\ H_1)}{P(data|\ H_0)}\right)\left(\frac{P(H_1)}{P(H_0)}\right).$$

Since, with LOD=3 the first term on the right-hand side is about 1,000:1 and the second term is about 1:50, the posterior odds against linkage are equal to 0.05. Since a multiple gene effect cannot be ruled out for a complex trait, he added one more hypothesis, the "other" hypothesis, H_2. The prior probabilities are given as follows, with h being the prior probability of H_2.

Hypothesis	Prior probability
H_0: Single gene, no linkage	0.98(1-h)
H_1: Single gene, linkage	0.02(1-h)
H_2: "Other"	h

The posterior odds are equal to,

$$Q = \frac{P(H_1|\ data)}{P(\overline{H_1}|\ data)} = \left(\frac{P(data|\ H_1)}{P(data|\ H_0)}\right)\left(\frac{P(H_1)}{P(H_0) + \left[\frac{P(data|\ H_2)}{P(data|\ H_0)}\right]P(H_2)}\right),$$

where $\overline{H_1}$ means "not H_1." Let $R = \frac{P(data|\ H_2)}{P(data|\ H_0)}$, then,

$$Q = \frac{0.02(1-h)}{0.98(1-h) + hR}.$$

If the disease in the family has an prior chance of 90% of being due to the "other" mode of inheritance, then the critical value has to increase from 3 to 4 in order to retain the posterior odds at 0.05.

The above Ott's Baysian argument applies to a single marker situation. When we do a genome-wide search, we use many markers to do multiple tests. While the overall

false positive rate increases due to the nature of multiple tests, Ott (1991) indicated that, because the prior probability also increase with the number of markers, the critical limit could remain at 3. He also reported that Lander and Botstein (1989) investigated the problem and concluded that for the human genome, in order to keep the overall significance level at 5%, the appropriate critical level of LOD is between 2 and 3. Later, Lander et al. (Lander and Schork, 1994; Lander and Kruglyak, 1995) strongly advocated higher LOD thresholds for genome-wide detection, results based on dense markers. This author have not tried to verify their "dense markers" approach hence will not include their results here, but would like to point out that for different methods the distribution of LOD are different. Thus, different thresholds should be applied to different methods as Lander and Schork, and Lander and Kruglyak did.

In practice, it seems that most researchers forgot the other part of Morton's suggestion, that is, when LOD < -2, the marker should be excluded. This author has not seen a paper published using the method of exclusion.

Clerget-Darpoux, Bonaiti-Pellié, and Hochez (1986) studied the effects of misspecified genetic parameters in LOD score analysis. They reported that "the power of the linkage test is sensitive to the degree of dominance, and slightly to the penetrance, but not to the gene frequency. In contrast, the estimation of the recombination fraction may be strongly affected by an error on any genetic parameter." MacLean et al. (1993) proposed a "MOD" method which is similar to LOD but it not only maximizing LOD over recombination fraction but also maximizing over inheritance mode.

2.8 Heterogeneity and Homogeneity

There are two types of heterogeneity linkage analysis. One is allelic and the other is nonallelic; the latter is also referred to as locus heterogeneity. "With allelic heterogeneity, individuals differ from each other by having different alleles at the same

locus reponsible for the disease; in nonallelic heterogeneity, however, the disease is caused by different loci" (Ott, 1991). We will discuss only nonallelic heterogeneity in this section because there are more than one recombination fraction that can be detected by linkage analysis.

There are several methods to test for homogeneity. These methods include a method proposed by Morton (1956), and a method by Smith (1963). Morton's method to test homogeneity is to test whether all the families have the same recombination fraction or have different recombination fractions. He proposed a simple statistic,

$$2\ln(10) \times [\sum_{i=1}^{k} Z_i(\hat{\theta}_i) - Z(\hat{\theta})], \tag{2.9}$$

where $\hat{\theta}_i$ is the recombination fraction that maximize the LOD score of the ith family, and $Z_i(\hat{\theta}_i)$ denotes the maximized LOD score, $i = 1$ to k. The $Z(\hat{\theta})$ is the total LOD-score maximum that occurs at a value of $\hat{\theta}$ for all families combined. This statistic is assumed to have a χ^2 distribution with k - 1 df.

Smith (1963) assumed that there were two groups, a linked group and an unlinked group. The linked group is the group of families that have a disease gene at a locus linked with the markers. The members of unlinked group either have no gene that caused the disease or have a gene at a locus that is unlinked to the markers. Smith proposed a test for testing whether all families in the study are from the linked group. If ε is the true but unknown proportion of families belonging to the linked group with recombination fraction θ, let $\hat{\theta}$ and $\hat{\varepsilon}$ be the MLE of ε and θ. Under H_1, the LOD score of the ith family is given as $Z_i(\varepsilon, \theta) = \log[\varepsilon L_i(\theta) + (1 - \varepsilon) L_i(\frac{1}{2})]$. The total LOD score was equal to $Z(\varepsilon, \theta) = \sum Z_i(\varepsilon, \theta)$. He assessed the significance of nonallelic heterogeneity by the statistic

$$2\ln(10) \times [Z(\hat{\varepsilon}, \hat{\theta}|H_1) - Z(\varepsilon = 1, \hat{\theta}|H_0)]. \tag{2.10}$$

If H_0 is true, because $\varepsilon = 1$ is on the boundary of parameter space, the statistic follows a 50:50 mixture distribution of a chi-square distribution with 1 df and a mass=1 at 0, not a chi-square distribution with 1 df as originally suggested.

Ott (1983) compared Smith's method, which he called the A-test, and Morton's method, which he called the PS-test (M-test in his 1991 book). For a mixed situation (families with or without linkages between gene locus and marker locus), Ott compared the M-test with the A-test and found that the A-test was generally superior. However, since he erred on the distribution of the test statistics (2.10), his conclusion might need modification.

Using IBD scores, Chakravart, Badner, and Li (1987) proposed a method to test linkage and homogeneity using IBD scores in the case of an autosomal recessive gene. The test was basically a goodness-of-fit test. They concluded that while the power of their method for detecting linkage from sib-pair data is excellent, that for detecting the heterogeneity of linkage is not. Proceedural details are as follows.

For an autosomal recessive disease, the genotype of unaffected parents in multiplex families are Dd x Dd, where D and d are the normal and disease alleles, respectively. Among the affected sib-pair, let the probabilities of marker IBD =2, 1, and 0, be,

$$P_2 = Pr(IBD = 2) = y^2, \tag{2.11}$$

$$P_1 = Pr(IBD = 1) = 2xy, \tag{2.12}$$

$$P_0 = Pr(IBD = 0) = x^2, \tag{2.13}$$

where $x = 2\theta(1-\theta) \leq 0.5$ and $y = 1 - x$. The maximum likelihood estimator (MLE) for x is

$$\hat{x} = \frac{n_1 + 2n_0}{2n}, \tag{2.14}$$

where n_i are the numbers of affected sib-pair sharing i markers IBD, and $n = \sum n_i$. Because true value should not be greater than 0.5, they proposed a new estimator,

$\hat{u}$, where,

$$\hat{u} = \begin{cases} \hat{x} & \text{if } \hat{x} \leq 0.5 \\ 0.5 & \text{if } \hat{x} > 0.5. \end{cases}$$

This new estimator has a smaller variance under both null and alternative hypotheses than $\hat{x}$'s. The recombination value may be estimated by

$$\hat{\theta} = \frac{1 - \sqrt{1 - 2\hat{u}}}{2}.$$

Under genetic heterogeneity, the IDB score distribution will be a mixture of two binomial distributions under $\theta < 0.5$ and $\theta = 0.5$ in the proportions $c : 1 - c$. Thus the P_i's in (2.11) through (2.13) become

$$P_2 = c(1 - x)^2 + (1 - c)/4, \tag{2.15}$$

$$P_1 = 2cx(1 - x) + (1 - c)/2, \tag{2.16}$$

$$P_0 = cx^2 + (1 - c)/4. \tag{2.17}$$

Solving the above equatons , they got $c = \frac{P_2 - P_0}{P_2 - P_1 + P_0}$. Replacing P_i with its MLE, they obtained the MLE for c, and x under heterogeneity

$$\widehat{c} = \frac{(n_2 - n_0)^2}{n(n_2 - n_1 + n_0)} \tag{2.18}$$

$$\widehat{x_h} = \frac{n_2 - 2n_0}{2(n_2 - n_0)} \tag{2.19}$$

Then, they proposed a two stage method to test linkage and heterogeneity. First they tests linkage, if the null hypothesis of no linkage was rejected then they conclude heterogeneity. In the first stage, the statistic

$$T = 2\sqrt{(2n)}(\hat{x} - \frac{1}{2})$$

was used and the authors claimed that T had asymptotic normal distribution. They did not explain why not using $\hat{u}$.

The statistic

$$G = \frac{n_2^2}{n(1-x)^2} + \frac{n_1^2}{2nx(1-x)} + \frac{n_0^2}{nx^2} - n \tag{2.20}$$

is used to test heterogeneity. The authors claimed G is asymptotically distributed as a χ^2 variable with df 1. However, they did not prove it nor point out what value should be used for x; use $\widehat{x}$ or $\widehat{x_h}$. Monte-Carlo simulations were used to evaluate the power.

2.9 Polygenes

There are many traits in a population that have more variation and can not be categorized into distinct classes easily (Klug and Cummings, 1997). Traits exhibiting continuous variation may be controlled by two or more genes. Such traits are said to exhibit continuous or quantitative variation and are examples of **polygenic inheritance**. The hypothesis suggesting a large number of factors or genes were responsible for continuous phenotype are called the **multiple-factor** or **multiple-gene hypothesis**. These genes have a special name, **polygenes** (Griffiths et al., 1993). Klug and Cummings (1997, p. 95) summarized some major characteristics of multiple-factor hypothesis:

1. Characters controlled by multiple-factors can usually be quantified by measuring, weighing, counting, etc.

2. Two or more pairs of genes, located throughout the genome, account for the hereditary influence on the phenotype in an additive way. Because many genes may be involved, inheritance of this type is often called *polygenic*.

3. Each gene locus may be occupied by either an additive allele, which contributes a set amount to the phenotype, or by a nonadditive allele, which does not contribute quantitatively to the phenotype.

4. The total effect of each additive alleles at each locus, while small, it approximately equivalent to all other additive alleles at other gene sites.

5. Together, the genes controlling a single character produce substantial phenotypic variation.

6. Analysis of polygenic traits requires the study of large numbers of progeny from a population of organisms.

If there are more than two genes controlling the phenotype, then the number of genes involved and the effects of those genes are important to know. To address this, Tan and Chang (1972) proposed a method for estimating the number of genes for self-fertilized populations assuming that there were only two alleles for each gene and the effect of each gene is the same. This work was further expanded on by Tan and D'Angelo (1979) to estimate the numbers and effects of major genes and polygenes assuming that all major genes have the same effect and all polygenes have the same effect. Because their work were developed for self-fertilized population, we will not cover the details.

2.10 Two-stage Genome Search

Using the DNA-level genetic markers technology proposed by Botstein et al. (1980), Lander and Botstein (1986) concluded that the affected-sib-pair method can be used for a genome search for disease genes. A two-stage design, first a screening search to eliminate nonviable marker loci, and then an intensive search to identify gene location is an intuitive design. Although this design has already been used in

genome-wide searches, such as those by Davies et al. (1994) and Luo et al. (1995), no optimality of their designs was discussed. Thus, there were no guidlines on how to space the markers or how to allocate the available ASP in each stage. While Elston (1992, 1994) studied the "optimal" two-stage design and concluded that two-stage designs are more efficient than one-stage designs, he did not consider the statistical complexity of the multiple test nor he considered the interval search nature in genome-wide linkage detection or the resource constraints. In another paper, Brown et al. (1994) studied the multiple-stage approach for genome search using affected-pedigree-member, by simulation. Since Brown et al. used only simulation and only one pedigree was considered, it is not obvious how to apply their results to the ASP method. Darvasi and Soller (1994) have studied the optimal spacing of genetic markers for the QTL trait without considering the two-stage approach. Holmans and Craddock (1997) conducted a simulation studied on the efficient strategies for genome scanning using maximum-likelihood affected-sib-pair analysis. The situation they considered are: a 200 affected sib-pairs sample, four different sample allocation strategies, and five grid-tightening strategies. The risk ratio of sibs are equal to 2 or 3, each marker locus have five equi-frequent alleles, and there are five possible location of genes. Since their studies were simulation studies, it is not clear how to generalize their results. Furtheremore, they did not consider different sample size, nor optimal strategies under resource constrains.

The main goals of this dissertation is to answer the two stage design question as they apply to rare autosomal recessive and dominant diseases with a dichotomous phenotype (affected or not affected).

CHAPTER 3
TWO-STAGE GENOME SEARCH FOR SIMPLE MENDELIAN DISEASE

In a two-stage search method, the first stage uses part of the ASPs with a wide spread markers in the genome. The rest of the resource is to be used on those promising markers identified in the first stage. In this study, only the least favorable configuration, that gene that lies in the middle of two adjacent markers, will be used for constructing the designs.

There are three major statistical problems involved in deriving analytic solutions are summarized as follows:

1. Because all the markers on the same chromosome are linked, the IBD scores are not independent. In the case of a genome search, there are many markers spread along the chromosomes. We have to handle a high-dimensional joint distribution.

2. In the first stage, a number of loci will be chosen for the more intensive linkage analysis in the second stage. The number and position of these loci that pass to the second stage are random. This makes calculation of exact distribution in the second stage extremely difficult.

3. Since in the first stage only a small number of ASPs and a large number of markers are used, there will be many ties in the IBD scores. The asymptotical solutions using a continuous normal distribution cannot handle these ties.

In this dissertation, Problem 3 was handled by a multinormial distribution. We used independent model as an approximation for Problem 1 and 2, and checked the results with simulation.

3.1 Assumptions

1. There is one and only one disease gene that increases the probability of an individual being affected. However, the disease may have nongenetic causes.

2. Highly polymorphic equally spaced markers are available. When m markers are assigned in the first stage, their positions are $\frac{L}{2m}$ $\frac{3L}{2m}$, ..., $\frac{(2m-1)L}{2m}$, where L, the length of the genome, is equal to 3300 cM (Lewin, 1990).

3. For a rare autosomal recessive disease, the parents' disease genotypes are D_1d and D_2d; and for a rare dominant disease, the parents' disease genotypes are D_1d and D_2D_3, where d is the disease gene.

4. In the same family, the probability that the disease of one affected sib is caused by a gene and the other is not is negligible.

5. The cost of typing alleles is a constant, i.e., the cost of typing k markers from one person is the same as typing one marker from k persons. This assumption of cost ratio can be relaxed to suit practical situations, but the numerical result would need to be re-calculated.

6. We use the Two-alleles statistic (Day and Simons, 1976), i.e., let X_{ij}=1 if ith marker IBD score = 2 for the jth sib-pair, and X_{ij}=0 otherwise. In our analytic approach, the X_{ij} are assumed to be independent random variables, except the two X_{ij} adjacent to the gene.

3.2 Two-Stage Procedure

Suppose there are n ASPs and there are enough resources to type N marker loci. Three numbers need to be determined in a two-stage design: n_1 and m, the number of ASPs and the number of markers to be used in the first stage (Stage I), and r, the number of markers to be used in the second stage (Stage II). The markers chosen for the second stage are based on the statistic used by Day and Simons,

$$ {}_1S_i = \sum_{j=1}^{n_1} X_{ij}, \tag{3.1} $$

in the first stage, where X_{ij} is defined in Assumption 6. Ideally, the r markers with the highest scores are chosen for Stage II. However, in the event of ties, more than r of them may have to be chosen. Thus, R, the actual number of markers used in Stage II, is a random variable. The formal definition of R is: $R \geq r$, but if the marker(s) with the lowest score in this group (markers for stage II study) is (are) taken away, then the total number of remaining markers is smaller than r.

In Stage II, R markers on N_2 ASPs are to be typed, where N_2 is the largest number of sibs that can be used subject to the resource constraint. Since R is a random variable, N_2 is also a random variable. Without loss of generality, let the R markers be 0,1, ..., $R-1$, and the sib-pairs in the Stage II be $n_1+1, n_1+2, ..., n_1+N_2$. Thus, N_2 is the largest x such that $mn_1 + Rx \leq N$. Define

$$ {}_2S_i = \sum_{j=n_1+1}^{n_1+N_2} X_{ij}. \tag{3.2} $$

Then the marker with the uniquely highest ${}_2S_i$ is claimed to be the locus nearest the disease gene. If two adjacent markers have the same highest score, then the gene is claimed to lie between them; otherwise, the gene location is undetermined. In this case, the two markers are called a marker group.

Once the location, I, has been chosen, say $I = i$, the next step is to check whether we can claim linkage. Let $t_{\alpha,n_2,R}$ be the $100(1 - \alpha)$ percentile of the unique maximum of R binomial $(n_2, 0.25)$ random variables. If ${}_2S_I > t_{\alpha,n_2,R}$, then we claim there is a linkage at loci i at significance level α.

3.3 Probability of Allocating the Correct Marker

3.3.1 Analytic Approach

To make the analytic solution tractable, Assumption (6) is made on the distributions of X_{ij}. They are later compared with a more realistic situation by simulation.

Let markers be numbered from 0 to $m-1$. Without loss of generality, if the gene is at the end of the chromosome, we let the nearest marker be marker 0; and if the gene is between two markers, we let it be located in the middle of marker 0 and 1. If we assume that gene location is uniformly distributed along the genome, the probability of the gene at the end is $1/m$.

Let ${}_1S_{(r)}$ be the rth largest statistic in the set $\{{}_1S_i\}$ in the first stage. The probability of finding the gene by the two-stage method is,

$$
\begin{aligned}
& P(\text{The marker closest to the gene is found}) \\
= \; & P(\text{Gene is at the end of the chromosome, marker 0 is chosen in Stage II}) \quad (3.3) \\
& + P(\text{Gene is not at the end, either marker 0 or 1 is chosen in Stage II}). \quad (3.4)
\end{aligned}
$$

Eq. (3.3) can be written as a sum of the probabilities of three exclusive events; Eq. (3.5) through Eq. (3.7).

$$
\begin{aligned}
& \frac{1}{m} P(\text{Marker 0 is chosen at the end of Stage II} \mid \text{gene is at the end}) \\
= \; & \frac{1}{m} P({}_1S_0 \text{ passes Stage I and } {}_2S_0 \text{ is the highest in Stage II})
\end{aligned}
$$

$$= \frac{1}{m}\sum_{k=0}^{n_1} P({}_1S_0 \text{ passes Stage I, } {}_2S_0 \text{ is the highest in Stage II, } {}_1S_{(r-1)} = k)$$

$$= \frac{1}{m}\sum_{k=0}^{n_1} \{P({}_1S_0 \geq {}_1S_{(r-1)},\ {}_1S_{(r-1)} = k,\ {}_2S_0 \text{ is the highest in Stage II}) \quad (3.5)$$

$$+ P(k = {}_1S_{(r-1)} \geq {}_1S_0 > {}_1S_{(r)},\ {}_2S_0 \text{ is the highest in Stage II}) \quad (3.6)$$

$$+ P(k = {}_1S_{(r-1)} \geq {}_1S_0 = {}_1S_{(r)},\ {}_2S_0 \text{ is the highest in Stage II})\}. \quad (3.7)$$

Eq. (3.5) is equal to,

$$P({}_1S_0 \geq {}_1S_{(r-1)},\ {}_1S_{(r-1)} = k, {}_2S_0 \text{ is the highest in Stage II})$$

(assuming independence)

$$= \sum_{i} \sum_{l} \sum_{i+l \geq (r-1),\ i < (r-1)} P({}_1S_0 \geq k, \quad i\ {}_1S_j's > k, \quad l\ {}_1S_j's = k,$$

$$(m-1-i-l)\ {}_1S_j's < k, \quad {}_2S_0 > \max_{t \neq 0}\ {}_2S_t)$$

$$= \sum_{i=0}^{r-2} \sum_{l=r-1-i}^{m-1-i} \left\{ P({}_1S_0 \geq k)\ \frac{m-1!}{i!\ l!\ (m-1-i-l)!} \right.$$

$$P({}_1S_j > k)^i \quad P({}_1S_j = k)^l \quad P({}_1S_j < k)^{m-1-i-l}$$

$$\left. \left[\sum_{t=1}^{N_2} P({}_2S_0 = t) P({}_2S_j < t)^{i+l} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{i+l+1} \right]_{Int}.$$

Eq. (3.6) is equal to,

$$P({}_1S_{(r-1)} = k > {}_1S_0 > {}_1S_{(r)},\ \text{gene is found})$$

$$= \sum_{i=0}^{k-2} P({}_1S_{(r-1)} = k > {}_1S_0 > {}_1S_{(r)} = i,\ {}_2S_0 \text{ is uniquely highest})$$

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-1)-(r-1)} P(k = {}_1S_{(r-1)} > {}_1S_1 > {}_1S_{(r)} = i, \quad (r-1-l)\ {}_1S_j's > k,$$

$$l\ {}_1S_j's = k,\quad h\ {}_1S_j's = i,\quad (m-r-h)\ {}_1S_j's < i,\ {}_2S_0 \text{ is uniquely highest})$$

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-1)-(r-1)} \left\{ P(k > {}_1S_0 > i)\ \frac{m-1!}{(r-1-l)!\ l!\ \ h!\ ((m-1)-(r-1)-h)!} \right.$$

$$P({}_1S_j > k)^{r-1-l}\ P({}_1S_j = k)^l\ P({}_1S_j = i)^h\ P({}_1S_j < i)^{m-1-(r-1)-l}$$

$$\left. \left[\sum_{t=1}^{N_2} P({}_2S_0 = t) P({}_2S_j < t)^{r-1} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r} \right]_{Int}.$$

Eq. (3.7) is equal to,

$$P({}_1S_{(r-1)} = k > {}_1S_0 = {}_1S_{(r)},\ \text{gene is found})$$

$$= \sum_{i=0}^{k-1} P({}_1S_{(r-1)} = k > {}_1S_0 = {}_1S_{(r)} = i,\ {}_2S_0 \text{ is uniquely highest})$$

$$= \sum_{i=0}^{k-1} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-1)-(r-1)} P(k = {}_1S_{r-1} > {}_1S_1 = {}_1S_r = i,\ (r-1-l)\ {}_1S_j's > k,$$

$$l\ {}_1S_j's = k,\ h\ {}_1S_j's = i,\ (m-r-h)\ {}_1S_j's < i,\ {}_2S_0 \text{ is uniquely highest})$$

$$= \sum_{i=0}^{k-1} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-1)-(r-1)} \{P(k > {}_1S_0 = i)\ \frac{m-1!}{(r-1-l)!\ l!\ \ h!\ ((m-1)-(r-1)-h)!}$$

$$P({}_1S_j > k)^{r-1-l}\ P({}_1S_j = k)^l\ P({}_1S_j = i)^h\ P({}_1S_j < i)^{m-1-(r-1)-l}$$

$$\left. \left[\sum_{t=1}^{N_2} P({}_2S_0 = t) P({}_2S_j < t)^{r-1+h} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r+h} \right]_{Int}.$$

Eq. (3.4) can be written as

$$\frac{m-1}{m} P(\text{Marker 0 or 1 is chosen at the end of Stage II} \mid \text{gene is not at the end})$$

$$= \frac{m-1}{m} \{P({}_1S_0 \text{ passes Stage I and } {}_1S_1 \text{ does not, } {}_2S_0 \text{ is the highest in Stage II}) \quad (3.8)$$

$+P({}_1S_1$ passes Stage I and ${}_1S_0$ does not, ${}_2S_1$ is the highest in Stage II) (3.9)

$+P({}_1S_0$ and ${}_1S_1$ pass Stage I and ${}_2S_0$ or ${}_2S_1$ is the highest in Stage II)$\}$. (3.10)

Since the gene is assumed to be in the middle of two markers, Eq. (3.8) is equal to Eq. (3.9) and they

$$= P({}_1S_0 \text{ passes Stage I and } {}_1S_1 \text{ does not, } {}_2S_0 \text{ is the highest in Stage II})$$

$$= \sum_{k=1}^{n_1} P({}_1S_0 \text{ passes Stage I and } {}_1S_1 \text{ does not, } {}_2S_0 \text{ is the highest in Stage II,}$$

$${}_1S_{(r-1)} = k)$$

$$= \sum_{k=1}^{n_1} \{P({}_1S_0 \geq {}_1S_{(r-1)} = k > {}_1S_1, {}_2S_0 \text{ is the highest in Stage II}) \quad (3.11)$$

$$+ P({}_1S_{(r-1)} = k > {}_1S_0 \geq {}_1S_{(r)}, {}_1S_0 > {}_1S_1, {}_2S_0 \text{ is the highest in Stage II}) \quad (3.12)$$

$$+ P({}_1S_{(r-1)} = k > {}_1S_0 = {}_1S_{(r)}, {}_1S_0 > {}_1S_1, {}_2S_0 \text{ is the highest in Stage II})\} \quad (3.13)$$

Eq. (3.11):

$$P({}_1S_0 \geq {}_1S_{(r-1)} = k > {}_1S_1, {}_2S_0 \text{ is the highest in Stage II})$$

$$= \sum_{i+l \geq (r-1),\ i < (r-1)} \sum_{l} P({}_1S_0 \geq k,\ i\ {}_1S_j's > k\ ,\ l\ {}_1S_j's = k,$$

$$(m-2-i-l)\ {}_1S_j's < k,\ {}_1S_1 < k\ {}_2S_0 > \max_{t \neq 0} {}_2S_t\ )$$

$$= \sum_{i=0}^{r-2} \sum_{l=r-1-i}^{m-2-i} \left\{ P({}_1S_0 \geq k, {}_1S_1 < k) \frac{m-2!}{i!\ l!\ (m-2-i-l)!} \right.$$

$$P({}_1S_j > k)^i\ P({}_1S_j = k)^l\ P({}_1S_j < k)^{m-2-i-l}$$

$$\left. \left[\sum_{t=1}^{N_2} P({}_2S_0 = t) P({}_2S_j < t)^{i+l} \right] \right\}$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{i+l+1} \right]_{Int}$$

Eq. (3.12) is equal to,

$$P(k = {}_1S_{(r-1)} > {}_1S_0 > {}_1S_{(r)} = i,\ {}_1S_0 > {}_1S_1,\ {}_2S_0 \text{ is the highest in Stage II})$$

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-2)-(r-1)} P(k = {}_1S_{(r-1)} > {}_1S_0 > {}_1S_{(r)} = i,\ {}_1S_0 > {}_1S_1,$$

$$(r-1-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k,\ h\ {}_1S_j{}'s = i,\ (m-r-1-h)\ {}_1S_j{}'s < i,$$

$${}_2S_0 > \max_{t \neq 0,1} {}_2S_t)$$

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-2)-(r-1)}$$

$$\left\{ P(k > {}_1S_0 > i, {}_1S_0 > {}_1S_1) \frac{m-2!}{(r-1-l)!\ l!\ h!\ ((m-2)-(r-1)-h)!} \right.$$

$$P({}_1S_j > k)^{r-1-l}\ P({}_1S_j = k)^l\ P({}_1S_j = i)^h\ P({}_1S_j < i)^{m-r-1-h}$$

$$\left. \left[\sum_{t=1}^{N_2} P({}_2S_0 = t) P({}_2S_j < t)^{r-1} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r} \right]_{Int}.$$

Eq. (3.13) is equal to,

$$\sum_{i=1}^{k-2} P(k = {}_1S_{(r-1)} > {}_1S_0 = {}_1S_{(r)} = i,\ {}_1S_0 > {}_1S_1,\ {}_2S_0 \text{ is the highest in Stage II})$$

$$= \sum_{i=1}^{k-2} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-2)-(r-1)} P(k = {}_1S_{(r-1)} > {}_1S_0 = {}_1S_{(r)} = i,\ {}_1S_0 > {}_1S_1,$$

$$(r-1-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k, h\ {}_1S_j{}'s = i,\ (m-r-1-h)\ {}_1S_j{}'s < i,$$

$${}_2S_0 > \max_{t \neq 0,1} {}_2S_t)$$

$$= \sum_{i=1}^{k-2} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-2)-(r-1)}$$

$$\left\{ P({}_1S_0 = i, {}_1S_1 < i) \frac{m-2!}{(r-1-l)!\ l!\ h!\ ((m-2)-(r-1)-h)!} \right.$$

$$P({}_1S_j > k)^{r-1-l}\, P({}_1S_j = k)^{l}\, P({}_1S_j = i)^{h}\, P({}_1S_j < i)^{m-r-1-h}$$

$$\left[\sum_{t=1}^{N_2} P({}_2S_0 = t) P({}_2S_j < t)^{r-1+h}\right]\Bigg\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r+h}\right]_{Int}.$$

Eq. (3.10) is equivalent to

$P({}_1S_0$ and ${}_1S_1$ pass Stage I, gene is found in Stage II)

$$= \sum_{k=1}^{n_1} \{P({}_1S_0 \geq {}_1S_{(r-2)} = k,\ {}_1S_1 \geq {}_1S_{(r-2)} = k, \text{ gene is found in Stage II}) \quad (3.14)$$

$$+ P({}_1S_0 \geq {}_1S_{(r-2)} = k > {}_1S_1 > {}_1S_{(r-1)}, \text{ gene is found in Stage II}) \quad (3.15)$$

$$+ P({}_1S_0 \geq {}_1S_{(r-2)} = k > {}_1S_1 = {}_1S_{(r-1)}, \text{ gene is found in Stage II}) \quad (3.16)$$

$$+ P({}_1S_1 \geq {}_1S_{(r-2)} = k > {}_1S_0 > {}_1S_{(r-1)}, \text{ gene is found in Stage II}) \quad (3.17)$$

$$+ P({}_1S_1 \geq {}_1S_{(r-2)} = k > {}_1S_0 = {}_1S_{(r-1)}, \text{ gene is found in Stage II}) \quad (3.18)$$

$$+ P({}_1S_{(r-2)} = k > {}_1S_0 > {}_1S_1 > {}_1S_{(r-1)}, \text{ gene is found in Stage II}) \quad (3.19)$$

$$+ P({}_1S_{(r-2)} = k > {}_1S_0 > {}_1S_1 = {}_1S_{(r-1)}, \text{ gene is found in Stage II}) \quad (3.20)$$

$$+ P({}_1S_{(r-2)} = k > {}_1S_1 > {}_1S_0 > {}_1S_{(r-1)}, \text{ gene is found in Stage II}) \quad (3.21)$$

$$+ P({}_1S_{(r-2)} = k > {}_1S_1 > {}_1S_0 = {}_1S_{(r-1)}, \text{ gene is found in Stage II}) \quad (3.22)$$

$$+ P({}_1S_{(r-2)} = k > {}_1S_1 = {}_1S_0 > {}_1S_{(r-1)}, \text{ gene is found in Stage II}) \quad (3.23)$$

$$+ P({}_1S_{(r-2)} = k > {}_1S_1 = {}_1S_0 = {}_1S_{(r-1)}, \text{ gene is found in Stage II}) \quad (3.24)$$

$$+ P({}_1S_{(r-1)} = k > {}_1S_1 = {}_1S_0 > {}_1S_{(r)}, \text{ gene is found in Stage II}) \quad (3.25)$$

$$+ P({}_1S_{(r-1)} = k > {}_1S_1 = {}_1S_0 = {}_1S_{(r)}, \text{ gene is found in Stage II})\}. \quad (3.26)$$

Again, because the gene is in the middle of two markers, Eq. (3.15) is equal to Eq. (3.17), (3.16) is equal to (3.18), (3.19) is equal to (3.21), and (3.20) is equal to (3.22).

Eq. (3.14) is equal to,

$$P({}_1S_0 \geq {}_1S_{(r-2)} = k, {}_1S_1 \geq {}_1S_{(r-2)} = k, {}_2S_0 \text{ is uniquely highest in Stage II})$$

$$+ \quad P({}_1S_0 \geq {}_1S_{(r-2)} = k, {}_1S_1 \geq {}_1S_{(r-2)} = k, {}_2S_1 \text{ is uniquely highest in Stage II})$$

$$+ \quad P({}_1S_0 \geq {}_1S_{(r-2)} = k, {}_1S_1 \geq {}_1S_{(r-2)} = k, {}_2S_0 = {}_2S_1 \text{is the highest})$$

$$= \quad \underbrace{P({}_1S_0 \geq {}_1S_{(r-2)} = k, {}_1S_1 \geq {}_1S_{(r-2)} = k, {}_2S_0 \text{ or } {}_2S_1 \text{ is uniquely highest in Stage II})}_{(A)}$$

$$+ \quad \underbrace{P({}_1S_0 \geq {}_1S_{(r-2)} = k, {}_1S_1 \geq {}_1S_{(r-2)} = k, {}_2S_0 = {}_2S_1 \text{is the highest})}_{(B)}$$

(A)

$$= \sum_{i} \sum_{l} P({}_1S_0 \geq k,\ {}_1S_1 \geq k,\ {}_1S_{(r-2)} = k, i\ {}_1S_j's > k,$$

$$l\ {}_1S_j's = k,\ (m-2-i-l)\ {}_1S_j's < k,\ {}_2S_0 \text{ or } {}_2S_1 \text{ is uniquely highest in Stage II})$$

$$= \sum_{i=0}^{r-3} \sum_{l=r-2-i}^{m-2-i} \left\{ P({}_1S_0 \geq k, {}_1S_1 \geq k) \frac{m-2!}{i!\,l!\ ((m-2)-i-l)!} \right.$$

$$P({}_1S_j > k)^i\ P({}_1S_j = k)^l\ P({}_1S_j < k)^{m-2-i-l}$$

$$\left. \left[P({}_2S_0 > \max_{t \neq 0,1} {}_2S_t, {}_2S_0 > {}_2S_1) + P({}_2S_1 > \max_{t \neq 0,1} {}_2S_t,\ {}_2S_1 > {}_2S_0) \right] \right\}$$

$$= \sum_{i=0}^{r-3} \sum_{l=r-2-i}^{m-2-i} \left\{ P({}_1S_0 \geq k, {}_1S_1 \geq k) \frac{m-2!}{i!\,l!\ ((m-2)-i-l)!} \right.$$

$$P({}_1S_j > k)^i\ P({}_1S_j = k)^l\ P({}_1S_j < k)^{m-2-i-l}$$

$$\left. 2 \left[\sum_{t=1}^{N_2} P({}_2S_0 = t, {}_2S_1 < t) P({}_2S_j < t)^{i+l} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r+l+2} \right]_{Int}.$$

(B)

$$= \sum_{i} \sum_{l} P({}_1S_0 \geq k,\ {}_1S_1 \geq k,\ {}_1S_{(r-2)} = k, i\ {}_1S_j's > k,$$

$l\ {}_1S_j's = k,\ (m-2-i-l)\ {}_1S_j's < k,\ {}_2S_0 = {}_2S_1$ is uniquely highest in Stage II)

$$= \sum_{i=0}^{r-3} \sum_{l=r-2-i}^{m-2-i} \left\{ P({}_1S_0 \geq k, {}_1S_1 \geq k) \frac{m-2!}{i!\ l!\ ((m-2)-i-l)!} \right.$$

$$P({}_1S_j > k)^i\ P({}_1S_j = k)^l\ P({}_1S_j < k)^{m-2-i-l}$$

$$\left. \left[P({}_2S_0 = {}_2S_1 > \max_{t \neq 0,1} {}_2S_t \right] \right\}$$

$$= \sum_{i=0}^{r-3} \sum_{l=r-2-i}^{m-2-i} \left\{ P({}_1S_0 \geq k, {}_1S_1 \geq k) \frac{m-2!}{i!\ l!\ ((m-2)-i-l)!} \right.$$

$$P({}_1S_j > k)^i\ P({}_1S_j = k)^l\ P({}_1S_j < k)^{m-2-i-l}$$

$$\left. \left[\sum_{t=1}^{N_2} P({}_2S_0 = t, {}_2S_1 = t) P({}_2S_j < t)^{r+l} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r+l+2} \right]_{Int}.$$

$(A) + (B)$

$$= \sum_{i=0}^{r-3} \sum_{l=r-2-i}^{m-2-i} \left\{ P({}_1S_0 \geq k, {}_1S_1 \geq k) \frac{m-2!}{i!\ l!\ ((m-2)-i-l)!} \right.$$

$$P({}_1S_j > k)^i\ P({}_1S_j = k)^l\ P({}_1S_j < k)^{m-2-i-l}$$

$$\left. \left[\sum_{t=1}^{N_2} [2P({}_2S_0 = t, {}_2S_1 < t) + P({}_2S_0 = t, {}_2S_1 = t)] P({}_2S_j < t)^{r+l} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r+l+2} \right]_{Int}.$$

Eq. (3.15) and (3.17) are equal to

$$P({}_1S_0 \geq {}_1S_{(r-2)} = k > {}_1S_1 > {}_1S_{(r-1)},\ \text{gene is found})$$

$$= \sum_{i=0}^{k-2} \{ P({}_1S_0 \geq {}_1S_{(r-2)} = k > {}_1S_1 > {}_1S_{(r-1)} = i, {}_2S_0 \text{ or } {}_2S_1$$

is uniquely highest in Stage II)

$+P({}_1S_0 \geq {}_1S_{(r-2)} = k > {}_1S_1 > {}_1S_{(r-1)} = i, {}_2S_0 = {}_2S_1$ is the highest in Stage II)

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-2)-(r-2)} \{P({}_1S_0 \geq k, k > {}_1S_1 > i, (r-1-l)\ {}_1S_j{}'s > k,$$

$l\ {}_1S_j{}'s = k, h\ {}_1S_j{}'s = i,\ (m-r-1-h)\ {}_1S_j{}'s < i, {}_2S_0$ or ${}_2S_1$

is uniquely highest in Stage II)

$+P({}_1S_0 \geq k, k > {}_1S_1 > i, (r-1-l)\ {}_1S_j{}'s > k,$

$l\ {}_1S_j{}'s = k, h\ {}_1S_j{}'s = i,\ (m-r-1-h)\ {}_1S_j{}'s < i, {}_2S_0 = {}_2S_1$

is the highest in Stage II)\}

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-2)-(r-2)}$$

$$\left\{ P({}_1S_0 \geq k, k > {}_1S_1 > i) \frac{m-2!}{(r-2-l)!\ l!\ h!\ ((m-2)-(r-2)-h)!} \right.$$

$$P({}_1S_j > k)^{r-2-l}\ P({}_1S_j = k)^l\ P({}_1S_j = i)^h\ P({}_1S_j < i)^{m-2-(r-2)-l}$$

$$\left. \left[\sum_{t=1}^{N_2} (2P({}_2S_0 = t, {}_2S_1 < t) + P({}_2S_0 = t, {}_2S_1 = t)) P({}_2S_j < t)^{r-2} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r} \right]_{Int}.$$

Eq. (3.16) and (3.18) are equal to

$P({}_1S_0 \geq {}_1S_{(r-2)} = k > {}_1S_1 = {}_1S_{(r-1)},$ gene is found)

$$= \sum_{i=0}^{k-1} \{P({}_1S_0 \geq {}_1S_{(r-2)} = k > {}_1S_1 = {}_1S_{(r-1)} = i, {}_2S_0 \text{ or } {}_2S_1$$

is uniquely highest in Stage II)

$+\ P({}_1S_0 \geq {}_1S_{(r-2)} = k > {}_1S_1 = {}_1S_{(r-1)} = i, {}_2S_0 = {}_2S_1$ is the highest)

$$= \sum_{i=0}^{k-1} \sum_{l=1}^{r-2} \sum_{h=1}^{(m-2)-(r-2)} \{P({}_1S_0 \geq k, {}_1S_1 > i, (r-2-l)\ {}_1S_j{}'s > k,$$

$l\ {}_1S_j{}'s = k, h\ {}_1S_j{}'s = i,\ (m-r-h)\ {}_1S_j{}'s < i, {}_2S_0$ or ${}_2S_1$

is uniquely highest in Stage II)

$$+ \quad P({}_1S_0 \geq k, {}_1S_1 >= i, (r-2-l)\ {}_1S_j{}'s > k,$$

$$l\ {}_1S_j{}'s = k, h\ {}_1S_j{}'s = i,\ (m-r-h)\ {}_1S_j{}'s < i, {}_2S_0 = {}_2S_1 \text{ is the highest})\}$$

$$= \sum_{i=0}^{k-1} \sum_{l=1}^{r-2} \sum_{h=1}^{(m-2)-(r-2)}$$

$$\left\{ P({}_1S_0 \geq k, {}_1S_1 > i)\ \frac{m-2!}{(r-2-l)!\ l!\ h!\ ((m-2)-(r-2)-h)!} \right.$$

$$P({}_1S_j > k)^{r-2-l}\ P({}_1S_j = k)^l\ P({}_1S_j = i)^h\ P({}_1S_j < i)^{m-2-(r-2)-l}$$

$$\left. \left[\sum_{t=1}^{N_2} (2P({}_2S_0 = t, {}_2S_1 < t) + P({}_2S_0 = t, {}_2S_1 = t)) P({}_2S_j < t)^{r-2+h} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r+h} \right]_{Int}.$$

Eq. (3.19) and (3.21) are equal to

$$P({}_1S_{(r-2)} = k > {}_1S_0 > {}_1S_1 > {}_1S_{(r-1)}, \text{ gene is found})$$

$$= \sum_{i=0}^{k-1} \{P({}_1S_{(r-2)} = k > {}_1S_0 > {}_1S_1 > {}_1S_{(r-1)} = i, {}_2S_0 \text{ or } {}_2S_1$$

is uniquely highest in Stage II)

$$+ \quad P({}_1S_{(r-2)} = k > {}_1S_0 > {}_1S_1 > {}_1S_{(r-1)} = i, {}_2S_0 = {}_2S_1 \text{is the highest})\}$$

$$= \sum_{i=0}^{k-3} \sum_{l=1}^{r-2} \sum_{h=1}^{(m-2)-(r-2)}$$

$$\{P(k > {}_1S_0 > {}_1S_1 > i, (r-2-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k,\ h\ {}_1S_j{}'s = i,$$

$$(m-r-h)\ {}_1S_j{}'s < i, {}_2S_0 \text{ or } {}_2S_1 \text{ is uniquely highest in Stage II})$$

$$+ \quad P(k > {}_1S_0 > {}_1S_1 >= i, (r-2-l)\ {}_1S_j{}'s > k,$$

$$l\ {}_1S_j{}'s = k, h\ {}_1S_j{}'s = i,\ (m-r-h)\ {}_1S_j{}'s < i, {}_2S_0 = {}_2S_1 \text{is the highest})\}$$

$$= \sum_{i=0}^{k-3} \sum_{l=1}^{r-2} \sum_{h=1}^{(m-2)-(r-2)}$$

$$\left\{ P(k > {}_1S_0 > {}_1S_1 > i)\ \frac{m-2!}{(r-2-l)!\ l!\ h!\ ((m-2)-(r-2)-h)!} \right.$$

$$P({}_1S_j > k)^{r-2-l}\ P({}_1S_j = k)^l\ P({}_1S_j = i)^h\ P({}_1S_j < i)^{m-2-(r-2)-h}$$

$$\left. \left[\sum_{t=1}^{N_2} (2P({}_2S_0 = t, {}_2S_1 < t) + P({}_2S_0 = t, {}_2S_1 = t)) P({}_2S_j < t)^{r-2} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r} \right]_{Int}.$$

Eq. (3.20) and (3.22) are equal to

$$P({}_1S_{(r-2)} = k > {}_1S_0 > {}_1S_1 = {}_1S_{(r-1)},\ \text{gene is found})$$

$$= \sum_{i=0}^{k-2} \{ P({}_1S_{(r-2)} = k > {}_1S_0 > {}_1S_1 = {}_1S_{(r-1)} = i, {}_2S_0 \text{ or } {}_2S_1$$

is uniquely highest in Stage II)

$$+ \quad P({}_1S_{(r-2)} = k > {}_1S_0 > {}_1S_1 = {}_1S_{(r-1)} = i, {}_2S_0 = {}_2S_1 \text{ is the highest in Stage II)} \}$$

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-2} \sum_{h=1}^{(m-2)-(r-2)}$$

$$\{ P(k > {}_1S_0 > i, {}_1S_1 = i, (r-2-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k,\ h\ {}_1S_j{}'s = i,$$

$$(m-r-h)\ {}_1S_j{}'s < i, {}_2S_0 \text{ or } {}_2S_1 \text{ is uniquely highest in Stage II)}$$

$$+ \quad P(k > {}_1S_0 > i, {}_1S_1 = i, (r-2-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k, h\ {}_1S_j{}'s = i,$$

$$(m-r-h)\ {}_1S_j{}'s < i, {}_2S_0 = {}_2S_1 \text{is the highest in Stage II)} \}$$

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-2} \sum_{h=1}^{(m-2)-(r-2)}$$

$$\{ P(k > {}_1S_0 > i, {}_1S_1 = i)\ \frac{m-2!}{(r-2-l)!\ l!\ \ h!\ ((m-2)-(r-2)-h)!}$$

$$P({}_1S_j > k)^{r-2-l}\ P({}_1S_j = k)^l\ P({}_1S_j = i)^h\ P({}_1S_j < i)^{m-2-(r-2)-h}$$

$$\left. \left[\sum_{t=1}^{N_2} (2P({}_2S_0 = t, {}_2S_1 < t) + P({}_2S_0 = t, {}_2S_1 = t)) \times P({}_2S_j < t)^{r-2+h} \right] \right\},$$

where $N_2 = \left[\frac{N - n_1 m}{r + h}\right]_{Int}$.

Eq. (3.23) is equal to

$$P({}_1S_{(r-2)} = k > {}_1S_0 = {}_1S_1 > {}_1S_{(r-1)}, \text{ gene is found})$$

$$= \sum_{i=0}^{k-2} \{P({}_1S_{(r-2)} = k > {}_1S_0 = {}_1S_1 > {}_1S_{(r-1)} = i, {}_2S_0 \text{ or } {}_2S_1$$

is uniquely highest in Stage II)

$$+ \quad P({}_1S_{(r-2)} = k > {}_1S_0 = {}_1S_1 > {}_1S_{(r-1)} = i, {}_2S_0 = {}_2S_1 \text{ is the highest in Stage II)}$$

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-2} \sum_{h=1}^{(m-2)-(r-2)}$$

$$\{P(k > {}_1S_0 = {}_1S_1 > i, (r-2-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k,\ h\ {}_1S_j{}'s = i,$$

$$(m-r-h)\ {}_1S_j{}'s < i, {}_2S_0 \text{ or } {}_2S_1 \text{is uniquely highest in Stage II)}$$

$$+ \quad P(k > {}_1S_0 = {}_1S_1 > i, (r-2-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k, h\ {}_1S_j{}'s = i,$$

$$(m-r-h)\ {}_1S_j{}'s < i, {}_2S_0 = {}_2S_1 \text{is the highest in Stage II)}\}$$

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-2} \sum_{h=1}^{(m-2)-(r-2)}$$

$$\left\{P(k > {}_1S_0 = {}_1S_1 > i)\ \frac{m-2!}{(r-2-l)!\ l!\ h!\ ((m-2)-(r-2)-h)!}\right.$$

$$P({}_1S_j > k)^{r-2-l}\ P({}_1S_j = k)^l\ P({}_1S_j = i)^h\ P({}_1S_j < i)^{m-2-(r-2)-h}$$

$$\left.\left[\sum_{t=1}^{N_2} (2P({}_2S_0 = t, {}_2S_1 < t) + P({}_2S_0 = t, {}_2S_1 = t))P({}_2S_j < t)^{r-2}\right]\right\},$$

where $N_2 = \left[\frac{N - n_1 m}{r}\right]_{Int}$.

Eq. (3.24) is equal to

$$P({}_1S_{(r-2)} = k > {}_1S_0 = {}_1S_1 = {}_1S_{(r-1)}, \text{ gene is found})$$

$$= \sum_{i=0}^{k-1} \{P({}_1S_{(r-2)} = k > {}_1S_0 = {}_1S_1 = {}_1S_{(r-1)} = i, {}_2S_0 \text{ or } {}_2S_1$$

is uniquely highest in Stage II)

$$+ \quad P({}_1S_{(r-2)} = k > {}_1S_0 = {}_1S_1 = {}_1S_{(r-1)} = i, {}_2S_0 = {}_2S_1 \text{ is the highest in Stage II)}$$

$$= \sum_{i=0}^{k-1} \sum_{l=1}^{r-2} \sum_{h=1}^{(m-2)-(r-2)} \{P({}_1S_0 = {}_1S_1 = i, (r-2-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k,$$

$h\ {}_1S_j{}'s = i,\ (m-r-h)\ {}_1S_j{}'s < i, {}_2S_0$ or ${}_2S_1$ is uniquely highest in Stage II)

$$+ \quad P({}_1S_0 = {}_1S_1 = i, (r-2-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k,$$

$h\ {}_1S_j{}'s = i,\ (m-r-h)\ {}_1S_j{}'s < i, {}_2S_0 = {}_2S_1$ is the highest in Stage II)\}

$$= \sum_{i=0}^{k-1} \sum_{l=1}^{r-2} \sum_{h=1}^{(m-2)-(r-2)}$$

$$\left\{ P({}_1S_0 = {}_1S_1 = i) \frac{m-2!}{(r-2-l)!\ l!\ \ h!\ ((m-2)-(r-2)-h)!} \right.$$

$$P({}_1S_j > k)^{r-2-l}\ P({}_1S_j = k)^l\ P({}_1S_j = i)^h\ P({}_1S_j < i)^{m-2-(r-2)-h}$$

$$\left. \left[\sum_{t=1}^{N_2} (2P({}_2S_0 = t, {}_2S_1 < t) + P({}_2S_0 = t, {}_2S_1 = t)) \times P({}_2S_j < t)^{r-2+h} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r+h} \right]_{Int}.$$

Eq. (3.25) is equal to

$$P({}_1S_{(r-1)} = k > {}_1S_0 = {}_1S_1 > {}_1S_{(r)},\ \text{gene is found})$$

$$= \sum_{i=0}^{k-2} \{P({}_1S_{(r-1)} = k > {}_1S_0 = {}_1S_1 > {}_1S_{(r)} = i, {}_2S_0 \text{ or } {}_2S_1$$

is uniquely highest in Stage II)

$$+ \quad P({}_1S_{(r-1)} = k > {}_1S_0 = {}_1S_1 > {}_1S_{(r)} = i, {}_2S_0 = {}_2S_1 \text{ is the highest in Stage II)}$$

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-2)-(r-1)}$$

$$\{P(k > {}_1S_0 = {}_1S_1 > i, (r-1-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k,\ h\ {}_1S_j{}'s = i,$$

$$(m-r-h)\ {}_1S_j{}'s < i, {}_2S_0 \text{ or } {}_2S_1 \text{ is uniquely highest in Stage II})$$

$$+\ P(k > {}_1S_0 = {}_1S_1 > i, (r-1-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k,$$

$$h\ {}_1S_j{}'s = i,\ (m-r-h)\ {}_1S_j{}'s < i, {}_1S_1 = {}_2S_1 \text{is the highest in Stage II})\}$$

$$= \sum_{i=0}^{k-2} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-2)-(r-1)}$$

$$\left\{ P(k > {}_1S_0 = {}_1S_1 > i)\ \frac{m-2!}{(r-1-l)!\ l!\ \ h!\ ((m-2)-(r-1)-h)!} \right.$$

$$P({}_1S_j > k)^{r-1-l}\ P({}_1S_j = k)^l\ P({}_1S_j = i)^h\ P({}_1S_j < i)^{m-2-(r-1)-h}$$

$$\left. \left[\sum_{t=1}^{N_2} (2P({}_2S_0 = t, {}_2S_1 < t) + P({}_2S_0 = t, {}_2S_1 = t)) \times P({}_2S_j < t)^{r-1} \right] \right\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r+1} \right]_{Int}.$$

Eq. (3.26) is equal to

$$P({}_1S_{(r-1)} = k > {}_1S_0 = {}_1S_1 = {}_1S_{(r)}, \text{ gene is found})$$

$$= \sum_{i=0}^{k-1} \{P({}_1S_{(r-1)} = k > {}_1S_0 = {}_1S_1 = {}_1S_{(r)} = i, {}_2S_0 \text{ or } {}_2S_1$$

$$\text{is uniquely highest in Stage II})$$

$$+\ P({}_1S_{(r-1)} = k > {}_1S_0 = {}_1S_1 = {}_1S_{(r)} = i, {}_2S_0 = {}_2S_1 \text{ is the highest in Stage II})$$

$$= \sum_{i=0}^{k-1} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-2)-(r-1)} \{P({}_1S_0 = {}_1S_1 = i, (r-1-l)\ {}_1S_j{}'s > k,\ l\ {}_1S_j{}'s = k,$$

$$h\ {}_1S_j{}'s = i,\ (m-r-h)\ {}_1S_j{}'s < i, {}_2S_0 \text{ or } {}_2S_1 \text{ is uniquely highest in Stage II})$$

$$+\ P({}_1S_0 = {}_1S_1 = i, (r-1-l)\ {}_1S_j{}'s > k,$$

$$l\ {}_1S_j{}'s = k, h\ {}_1S_j{}'s = i,\ (m-r-h)\ {}_1S_j{}'s < i, {}_2S_0 = {}_2S_1 \text{ is the highest in Stage II})\}$$

Table 3.1. The joint distribution of IBD score of the locus 1 and locus 2.

	IBD of locus 2		
IBD of locus 1	2	1	0
2	$\frac{\Psi^2}{4}$	$\frac{\Psi(1-\Psi)}{2}$	$\frac{(1-\Psi)^2}{4}$
1	$\frac{\Psi(1-\Psi)}{2}$	$\frac{\Psi^2+(1-\Psi)^2}{2}$	$\frac{\Psi(1-\Psi)}{2}$
0	$\frac{(1-\Psi)^2}{4}$	$\frac{\Psi(1-\Psi)}{2}$	$\frac{\Psi^2}{4}$

Where $\theta^2 + (1-\theta)^2$ and θ is the recombination fraction between two loci.

$$= \sum_{i=0}^{k-1} \sum_{l=1}^{r-1} \sum_{h=1}^{(m-2)-(r-1)} \{P({}_1S_0 = {}_1S_1 = i) \frac{m-2!}{(r-1-l)!\ l!\ h!\ ((m-2)-(r-1)-h)!}$$

$$P({}_1S_j > k)^{r-1-l}\ P({}_1S_j = k)^l\ P({}_1S_j = i)^h\ P({}_1S_j < i)^{m-2-(r-1)-h}$$

$$\left[\sum_{t=1}^{N_2} (2P({}_2S_0 = t, {}_2S_1 < t) + P({}_2S_0 = t, {}_2S_1 = t)) \times P({}_2S_j < t)^{r-1+h} \right] \Bigg\},$$

$$\text{where } N_2 = \left[\frac{N - n_1 m}{r+1+h} \right]_{Int}.$$

The joint distribution of the IBD scores of two markers (or genes), P(IBD of the marker (gene) 1, IBD of marker (gene) 2) is given in Table 3.1 adapted from Table 2.3 where $\Psi=\theta^2 + (1-\theta)^2$ and θ is the recombination fraction between two loci.

From Table 3.1, we can deduce the joint distribution of X_{ij} and $X_{i'j}$ for $i \neq i'$. The joint distribution between X_{ij} and $X_{i'j}$ is shown in Table 3.2, with Ψ defined as the Ψ in Table 3.1.

If a marker is not linked with the disease gene, then the probabilities of getting an IBD score of the marker equal to 2, 1, and 0 are 0.25, 0.5, and 0.25, respectively. Thus, X_{ij} has a Bernoulli distribution with parameter of 0.25. The distribution of ${}_lS_i$ under the null hypotheses is Binomial(n_l, 0.25) for l=1, 2. Let ε be the probability that the disease is caused by the gene, and θ be the distance between marker 0 and the gene. For equations 3.5-3.7, where the recessive gene is at the end, X_{0j} is distributed

as Bernoulli($\varepsilon\ \Psi^2+(1-\varepsilon)\ 0.25$) and ${}_lS_0$ as Binomial(n_l, $\varepsilon\ \Psi^2+(1-\varepsilon)\ 0.25$). Similarly, for a dominant disease gene, X_{0j} is distributed as Bernoulli($\varepsilon\ \Psi/2+(1-\varepsilon)\ 0.25$) and ${}_lS_i$ as Binomial(n_l, $\varepsilon\ \Psi/2+(1-\varepsilon)\ 0.25$).

For Eq. 3.11-3.13, we need to find the joint distribution P(${}_lS_0$=i, ${}_lS_1$=j), i=0,1,...,n_l, j=0,1,...,n_l, l=1,2. Let n_{abl} be the number of (X_{0j}, X_{1j}) = (a,b) in stage l. Then (n_{00l}, n_{01l}, n_{10l}, n_{11l}) has a multinomial distribution (n_l, p_{00}, p_{01}, p_{10}, p_{11}). Thus, the joint distribution

$$P({}_lS_0=\ i,\ {}_lS_1=\ j) \tag{3.27}$$

$$= P(n_{10l}\ +\ n_{11l}{=}i,\ n_{01l}\ +\ n_{11l}{=}j) \tag{3.28}$$

$$= \sum_{k\ =\ \max(0,\ i\ +\ j\ -\ n)}^{\min(i,j\ )} P(n_{11l}{=}k,\ n_{10l}{=}i-k,\ n_{01l}{=}j-k,$$

$$n_{00l}{=}n-i-j+n_{11l}), \tag{3.29}$$

can be computed once p_{00}, p_{01}, p_{10}, and p_{11} are specified.

Although we assume the gene is in the middle of the two adjacent markers, the formulae are derived for the gene anywhere in between. Let the recombination fraction between the gene and marker 0 be θ_0 and between the gene and marker 1 be θ_1. Let θ_2 be the recombination fraction between two markers. The joint distribution of X_{0j} and X_{1j}, for $x=0,1, y=0,1$, is,

Table 3.2. The joint distribution of X_{ij} and $X_{i'j}$.

	$X_{i'j}$	
X_{ij}	1	0
1	$\frac{\Psi^2}{4}$	$\frac{(1-\Psi^2)}{4}$
0	$\frac{(1-\Psi^2)}{4}$	$\frac{\Psi^2+2}{4}$

$$P(X_{0j} = x, X_{1j} = y) \tag{3.30}$$

$$= P(X_{0j} = x, X_{1j} = y \mid \text{gene IBD=2})P(\text{gene IBD=2})$$
$$+ P(X_{0j} = x, X_{1j} = y \mid \text{gene IBD=1})P(\text{gene IBD=1})$$
$$+ P(X_{0j} = x, X_{1j} = y \mid \text{gene elsewhere})P(\text{gene elsewhere}) \tag{3.31}$$

$$= P(X_{0j} = x \mid X_{1j} = y \text{ , gene IBD=2})P(X_{1j} = y \mid \text{gene IBD=2})$$
$$P(\text{gene IBD=2})$$
$$+ P(X_{0j} = x \mid X_{1j} = y, \text{gene IBD=1})P(X_{1j} = y \mid \text{gene IBD=1})$$
$$P(\text{gene IBD=1})$$
$$+ P(X_{0j} = x \mid X_{1j} = y, \text{gene elsewhere})$$
$$P(X_{1j} = y \mid \text{gene elsewhere})P(\text{ gene elsewhere}). \tag{3.32}$$

Since an individual has to have two recessive disease genes in order to be affected, Eq. (3.32) becomes

$$= P(X_{0j} = x \mid X_{1j} = y, \text{gene IBD=2})P(X_{1j} = y \mid \text{gene IBD=2})\varepsilon$$
$$+ P(X_{0j} = x \mid X_{1j} = y, \text{gene elsewhere})P(X_{1j} = y \mid \text{gene elsewhere})(1 - \varepsilon).$$

For a dominant disease, if an ASP is caused by a gene, then both sibs must at least share the disease gene and there is a 50-50 chance they share the other allele. Thus, Eq. (3.32) becomes

$$= P(X_{0j} = x \mid X_{1j} = y, \text{gene IBD=2})P(X_{1j} = y \mid \text{gene IBD=2})(\varepsilon/2)$$
$$+ P(X_{0j} = x \mid X_{1j} = y, \text{gene IBD=1})P(X_{1j} = y \mid \text{gene IBD=1})(\varepsilon/2)$$
$$+ P(X_{0j} = x \mid X_{1j} = y, \text{gene elsewhere})$$
$$P(X_{1j} = y \mid \text{gene elsewhere})(1 - \varepsilon). \tag{3.33}$$

Let $\Psi_l = \theta_l^2 + (1 - \theta_l)^2$, l=0,1,2, $p_{xy} = P(X_{0j} = x \mid X_{1j} = y)$, $x = 0,1$, $y = 0,1$. Then

apply Table 3.2 for a recessive disease we have

$$p_{00} = (1-\Psi_0^2)(1-\Psi_1^2)\,\varepsilon + \frac{\Psi_2^2+2}{4}(1-\varepsilon), \tag{3.34}$$

$$p_{10} = \Psi_0^2(1-\Psi_1^2)\,\varepsilon + \frac{1-\Psi_2^2}{4}(1-\varepsilon), \tag{3.35}$$

$$p_{01} = (1-\Psi_0^2)\,\Psi_1^2\,\varepsilon + \frac{1-\Psi_2^2}{4}(1-\varepsilon), \tag{3.36}$$

$$p_{11} = \Psi_0^2\,\Psi_1^2\,\varepsilon + \frac{\Psi_2^2}{4}(1-\varepsilon). \tag{3.37}$$

For a dominant disease,

$$p_{00} = (1-\Psi_0^2)(1-\Psi_1^2)\,\frac{\varepsilon}{2} + (1-\Psi_0+\Psi_0^2)(1-\Psi_1+\Psi_1^2)\,\frac{\varepsilon}{2} + \frac{\Psi_2^2+2}{4}(1-\varepsilon), \tag{3.38}$$

$$p_{10} = \Psi_0^2(1-\Psi_1^2)\,\frac{\varepsilon}{2} + \Psi_0(1-\Psi_0)(1-\Psi_1+\Psi_1^2)\,\frac{\varepsilon}{2} + \frac{1-\Psi_2^2}{4}(1-\varepsilon), \tag{3.39}$$

$$p_{01} = (1-\Psi_0^2)\,\Psi_1^2\,\varepsilon + (1-\Psi_0+\Psi_0^2)\Psi_1(1-\Psi_1)\,\frac{\varepsilon}{2} + \frac{(1-\Psi_2^2)}{4}(1-\varepsilon), \tag{3.40}$$

$$p_{11} = \Psi_0^2\,\Psi_1^2\,\frac{\varepsilon}{2} + \Psi_0(1-\Psi_0)\Psi_1(1-\Psi_1)\,\frac{\varepsilon}{2} + \frac{\Psi_2^2}{4}(1-\varepsilon). \tag{3.41}$$

Thus, P(the marker closest to the gene is found) can be calculated. In addition to analysis computation for Eq. 3.3 and 3.4, simulations were also done.

3.3.2 Simulation under More Realistic Assumptions

The assumption of independent X_{ij} violates the fact that some loci are linked. Simulations were done under a Markov chain model with the combinations of resource $N = 1000$, 2000, 5000, and 10,000, number of ASP $n = 10$, 25, 75, and 100, $\varepsilon = 0.25$, 0.5, 0.75, and 1, m from 50 to 350 with an increment 25 for two-stage design and 10 for one-stage design, r from 5 to m with an increment $(m-5)/10$, and n_1 from 5 to $n-5$ with increment 5.

The simulations were conducted as follows:

1. Reading in parameters, resource N, heterogeneity ε, number of ASP n, and m, n_1, and r of the design.

2. Given n_1, m, generate y from uniform(0,1). If $y \leq \varepsilon$ then generate gene location from uniform distribution (0,3300). Let locus l and locus $l+1$ be two adjacent loci, i.e. the gene location is between $\frac{2lL}{2m}$ and $\frac{(2l+1)L}{2m}$, where L is the total length of genome. This step determines which interval contains the gene and then:

   (a) For recessive disease, generate IBD scores at loci l and $l+1$ conditional on both sibs of ASP carrying two disease genes and the gene locus is in the middle of two markers. Haldane map function is used to convert map distance into recombination fraction θ.

   (b) For dominant disease, if $y \leq 0.5\varepsilon$, generate IBD scores at loci l and $l+1$ conditional on both sibs of ASP sharing two alleles at gene loci and the gene locus is in the middle of two markers. If $0.5\varepsilon \leq y \leq \varepsilon$, then generate IBD score at loci l and $l+1$ conditional on both sibs of ASP share one allele at gene loci and, again, the gene locus is in the middle of two markers.

   If $y > \epsilon$, let l =0, and generate IBD score at locus 0 from Bernoulli(0.25) as if there is no gene linked with marker 0.

3. For Markov chain model simulation, generate IBD score at locus i, i start from $l-1$ and decrease to 0, conditional on IBD score at locus $i+1$, and then generate IBD score at locus i conditional on IBD score at locus $i-1$ for $i = l+1, ..., m-1$. For independent model simulation, generate IBD scores independently from Trinomial(n_1, 0.25, 0.5,0.25). Conditional probability formulas are given by Table 2.3.

4. Convert IBD scores into statistics X_{ij}s.

5. Repeat step 2, 3, and 4 n_1 times and then calculated ${}_1S_i$.

6. Check ties, adjust R to include all the ties with the rth highest ${}_1S_i$ for Stage II, check whether ${}_1S_0$ passes Stage I. If yes, then calculate N_2 according to the resource constraint and go to the next step. If not, record the detection as a failure and go back to step 2.

7. The X_{ij}s in Stage II are generated the same way as those in Stage I, except only the chosen markers are used and repeat N_2 times.

8. Check whether ${}_2S_0$ or ${}_2S_1$ is the unique largest among ${}_2S_i$s in Stage II.

9. If marker 0 or 1 is chosen, it is a success; otherwise it is a failure. Go back to step 2 until enough simulation has been done.

The programs were compiled using GCC version 2.7.2 on a PC with Pentium 166 CPU and 48M RAM in an OS/2 environment. The random number generator for the simulation was adapted from Press (1992).

3.3.3 Results

Based on analytic computation and simulation, the designs with the highest probability of finding the right marker (power) were identified. These optimal designs are given in Table 3.3 for searching a recessive gene, and Table 3.4 for a dominant gene. In both tables, in the ASP column is reported the number of available affected-sib-pairs; the ε column shows the probability that the disease is caused by the gene, representing heterogeneity; the m columns, the number of marker loci used in the first stage; the r column, the proposed number of loci to be chosen in Stage I for Stage II study; the n_1 columns, the number of ASP used in first stage; the F2 column, the probability of locating the right marker by the best two-stage design obtained by analytic formula; the Indep columns and the Markov columns show the probabilities obtained by simulation with independent assumption and Markov chain assumption without combining first- and second-stage data; and the Comb. column shows the

simulated probability of the two-stage design with Markov chain model with first- and second-stage data combined. The F1 column gives the probability of the best one-stage design, i.e., with optimal m and n subject to $mn \leq N$ obtained by analytic formula. The last column, F2−F1, shows the increase in probability of two-stage design over one-stage. An asterisk is marked when the increase was over 0.35 for Table 3.3 (recessive) and 0.15 for Table 3.4 (dominant).

Table 3.3. Optimal resource allocation in two-stage genome search for rare recessive gene.

Resource N=1000														
Parameter		Two-stage design							One-stage design					Improv.
ASP	ε	m	r	n_1	F2	Indep	Markov	Comb.	m	n	F1	Indep	Markov	F2-F1
10	0.25	125	5	5	0.040	0.041	0.041	0.045	100	10	0.042	0.043	0.042	-0.001
10	0.50	175	5	5	0.149	0.152	0.146	0.167	100	10	0.121	0.126	0.116	0.027
10	0.75	175	5	5	0.366	0.369	0.352	0.412	100	10	0.269	0.268	0.272	0.098
10	1.00	175	5	5	0.662	0.662	0.631	0.723	100	10	0.481	0.484	0.477	0.181
25	0.25	125	5	5	0.092	0.095	0.090	0.093	50	20	0.059	0.062	0.059	0.034
25	0.50	125	5	5	0.324	0.339	0.316	0.324	70	14	0.127	0.126	0.130	0.197
25	0.75	125	5	5	0.642	0.654	0.630	0.642	100	10	0.269	0.269	0.268	0.374*
25	1.00	150	5	5	0.897	0.888	0.882	0.894	100	10	0.481	0.495	0.474	0.416*
50	0.25	100	5	5	0.126	0.130	0.135	0.135	50	20	0.059	0.061	0.060	0.068
50	0.50	125	5	5	0.388	0.385	0.372	0.377	70	14	0.127	0.129	0.120	0.260
50	0.75	125	5	5	0.697	0.701	0.680	0.685	100	10	0.269	0.277	0.272	0.428*
50	1.00	125	5	5	0.902	0.898	0.894	0.897	100	10	0.481	0.479	0.474	0.421*
75	0.25	100	5	5	0.135	0.141	0.134	0.136	50	20	0.059	0.059	0.062	0.077
75	0.50	100	5	5	0.400	0.404	0.401	0.404	70	14	0.127	0.125	0.133	0.273
75	0.75	125	5	5	0.697	0.707	0.689	0.696	100	10	0.269	0.278	0.269	0.429*
75	1.00	125	5	5	0.902	0.898	0.887	0.892	100	10	0.481	0.488	0.474	0.421*
100	0.25	100	5	5	0.137	0.138	0.136	0.138	50	20	0.059	0.059	0.061	0.078
100	0.50	100	5	5	0.402	0.419	0.387	0.393	70	14	0.127	0.129	0.125	0.274
100	0.75	125	5	5	0.697	0.703	0.685	0.691	100	10	0.269	0.268	0.268	0.429*
100	1.00	125	5	5	0.902	0.900	0.891	0.895	100	10	0.481	0.481	0.477	0.421*

Table 3.3–continued.

Resource N=2000														
Parameter		Two Stage Design							One-stage design					Improv.
ASP	ε	m	r	n_1	F2	Indep	Markov	Comb.	m	n	F1	Indep	Markov	F2-F1
10	0.25	125	5	5	0.040	0.042	0.041	0.044	200	10	0.041	0.043	0.044	-0.001
10	0.50	350	5	5	0.181	0.198	0.154	0.170	200	10	0.178	0.188	0.177	0.003
10	0.75	350	5	5	0.500	0.513	0.419	0.481	200	10	0.460	0.472	0.441	0.040
10	1.00	350	5	5	0.883	0.880	0.745	0.834	200	10	0.807	0.808	0.779	0.076
25	0.25	175	5	10	0.108	0.120	0.114	0.118	80	25	0.085	0.090	0.088	0.023
25	0.50	175	5	10	0.441	0.455	0.427	0.453	100	20	0.269	0.272	0.265	0.172
25	0.75	175	5	10	0.824	0.830	0.801	0.832	110	18	0.565	0.570	0.563	0.259
25	1.00	175	5	10	0.984	0.984	0.974	0.985	140	14	0.841	0.845	0.829	0.143
50	0.25	150	5	10	0.177	0.184	0.168	0.173	60	33	0.088	0.095	0.094	0.089
50	0.50	150	5	10	0.594	0.611	0.578	0.588	100	20	0.269	0.275	0.266	0.326
50	0.75	150	5	10	0.893	0.905	0.887	0.891	110	18	0.565	0.570	0.561	0.329
50	1.00	175	5	10	0.990	0.990	0.987	0.992	140	14	0.841	0.852	0.829	0.149
75	0.25	150	5	5	0.213	0.221	0.206	0.208	60	33	0.088	0.087	0.093	0.124
75	0.50	150	5	10	0.614	0.626	0.605	0.610	100	20	0.269	0.274	0.269	0.345
75	0.75	100	14	10	0.897	0.901	0.891	0.905	110	18	0.565	0.565	0.567	0.332
75	1.00	175	5	10	0.990	0.990	0.986	0.991	140	14	0.841	0.838	0.828	0.149
100	0.25	150	5	5	0.229	0.240	0.219	0.221	60	33	0.088	0.089	0.098	0.141
100	0.50	125	5	10	0.623	0.638	0.618	0.621	100	20	0.269	0.283	0.266	0.354*
100	0.75	100	14	10	0.897	0.903	0.884	0.899	110	18	0.565	0.579	0.559	0.332
100	1.00	175	5	10	0.990	0.990	0.988	0.991	140	14	0.841	0.848	0.837	0.149

Table 3.3–continued.

Resource N=5000														
Parameter		Two-stage design							One-stage design					Improv.
ASP	ε	m	r	n_1	F2	Indep	Markov	Comb.	m	n	F1	Indep	Markov	F2-F1
10	0.25	125	5	5	0.040	0.039	0.037	0.041	210	10	0.041	0.046	0.035	-0.001
10	0.50	350	5	5	0.181	0.202	0.152	0.171	350	10	0.197	0.212	0.175	-0.016
10	0.75	350	5	5	0.500	0.521	0.425	0.480	350	10	0.552	0.566	0.491	-0.052
10	1.00	350	5	5	0.883	0.882	0.742	0.830	350	10	0.929	0.933	0.838	-0.047
25	0.25	300	5	15	0.118	0.130	0.114	0.128	200	25	0.127	0.138	0.131	-0.009
25	0.50	350	73	5	0.578	0.615	0.516	0.572	200	25	0.555	0.577	0.534	0.022
25	0.75	350	73	5	0.951	0.956	0.888	0.930	200	25	0.926	0.938	0.911	0.025
25	1.00	350	73	5	1.000	1.000	0.986	0.994	350	14	0.989	0.992	0.943	0.011
50	0.25	250	29	10	0.270	0.291	0.257	0.278	100	50	0.193	0.198	0.197	0.077
50	0.50	250	29	10	0.860	0.881	0.825	0.853	130	38	0.640	0.658	0.632	0.221
50	0.75	225	27	15	0.995	0.997	0.986	0.995	170	29	0.937	0.951	0.932	0.058
50	1.00	200	5	20	1.000	1.000	0.999	1.000	350	14	0.989	0.992	0.943	0.011
75	0.25	200	24	10	0.367	0.395	0.354	0.366	100	50	0.193	0.194	0.191	0.174
75	0.50	200	24	15	0.915	0.928	0.893	0.907	130	38	0.640	0.661	0.634	0.275
75	0.75	150	19	20	0.998	0.998	0.995	0.997	170	29	0.937	0.948	0.929	0.061
75	1.00	125	5	30	1.000	1.000	1.000	1.000	350	14	0.989	0.992	0.941	0.011
100	0.25	175	22	15	0.418	0.441	0.405	0.424	100	50	0.193	0.200	0.193	0.225
100	0.50	150	19	20	0.929	0.938	0.921	0.934	130	38	0.640	0.659	0.637	0.289
100	0.75	125	5	35	0.999	0.997	0.997	0.998	170	29	0.937	0.949	0.928	0.061
100	1.00	100	5	35	1.000	1.000	1.000	1.000	350	14	0.989	0.991	0.944	0.011

Table 3.3–continued.

Resource N=10,000														
Parameter		Two-stage design							One-stage design					Improv.
ASP	ε	m	r	n_1	F2	Indep	Markov	Comb.	m	n	F1	Indep	Markov	F2-F1
10	0.25	125	5	5	0.040	0.042	0.042	0.044	210	10	0.041	0.043	0.046	-0.001
10	0.50	350	5	5	0.181	0.200	0.156	0.175	350	10	0.197	0.219	0.179	-0.016
10	0.75	350	5	5	0.500	0.515	0.422	0.474	350	10	0.552	0.570	0.493	-0.052
10	1.00	350	5	5	0.883	0.885	0.739	0.830	350	10	0.929	0.928	0.835	-0.047
25	0.25	300	5	15	0.118	0.127	0.114	0.129	350	25	0.132	0.148	0.129	-0.014
25	0.50	350	73	5	0.578	0.606	0.523	0.579	350	25	0.647	0.685	0.590	-0.069
25	0.75	350	73	5	0.951	0.958	0.900	0.930	350	25	0.974	0.982	0.934	-0.023
25	1.00	350	73	5	1.000	0.999	0.985	0.993	350	25	0.997	1.000	0.993	0.003
50	0.25	350	73	5	0.294	0.326	0.269	0.279	200	50	0.295	0.314	0.281	-0.000
50	0.50	325	101	10	0.907	0.927	0.852	0.902	200	50	0.894	0.912	0.872	0.013
50	0.75	325	101	10	0.999	1.000	0.990	0.996	350	28	0.985	0.992	0.958	0.014
50	1.00	225	5	20	1.000	1.000	0.999	1.000	350	28	0.997	1.000	0.994	0.003
75	0.25	275	59	15	0.449	0.488	0.430	0.467	140	71	0.361	0.381	0.356	0.088
75	0.50	275	59	15	0.976	0.980	0.953	0.969	160	62	0.918	0.937	0.906	0.059
75	0.75	250	5	35	1.000	1.000	0.996	1.000	350	28	0.985	0.992	0.956	0.014
75	1.00	125	5	30	1.000	1.000	1.000	1.000	350	28	0.997	1.000	0.996	0.003
100	0.25	250	29	25	0.560	0.600	0.527	0.561	100	100	0.386	0.408	0.386	0.174
100	0.50	225	27	30	0.990	0.993	0.977	0.990	160	62	0.918	0.934	0.912	0.073
100	0.75	150	5	40	1.000	1.000	0.999	1.000	350	28	0.985	0.991	0.957	0.015
100	1.00	100	5	35	1.000	1.000	1.000	1.000	350	28	0.997	1.000	0.996	0.003

Table 3.4. Optimal resource allocation in two-stage genome search for rare dominant gene.

Resource N=1000														
Parameter		Two-stage design							One-stage design					Improv.
ASP	ϵ	m	r	n_1	F2	Indep	Markov	Comb.	m	n	F1	Indep	Markov	F2-F1
10	0.25	50	5	5	0.028	0.029	0.027	0.030	50	10	0.028	0.026	0.029	0.000
10	0.50	50	5	5	0.036	0.034	0.037	0.037	50	10	0.037	0.035	0.039	-0.000
10	0.75	100	5	5	0.050	0.053	0.051	0.059	100	10	0.053	0.059	0.053	-0.003
10	1.00	175	5	5	0.082	0.084	0.082	0.092	100	10	0.082	0.078	0.080	-0.000
25	0.25	50	5	5	0.042	.043	0.039	0.040	50	20	0.036	0.036	0.041	0.005
25	0.50	75	5	10	0.063	0.064	0.066	0.067	50	20	0.053	0.054	0.053	0.009
25	0.75	125	5	5	0.119	0.122	0.121	0.124	50	20	0.075	0.076	0.074	0.044
25	1.00	125	5	5	0.201	0.205	0.197	0.204	50	20	0.103	0.096	0.106	0.098
50	0.25	50	5	10	0.053	0.053	0.052	0.053	50	20	0.036	0.037	0.039	0.016
50	0.50	100	5	5	0.091	0.093	0.088	0.089	50	20	0.053	0.055	0.058	0.037
50	0.75	100	5	5	0.168	0.168	0.162	0.162	50	20	0.075	0.078	0.070	0.092
50	1.00	100	5	5	0.270	0.264	0.258	0.262	50	20	0.103	0.104	0.102	0.166*
75	0.25	50	5	10	0.057	0.058	0.058	0.059	50	20	0.036	0.035	0.036	0.020
75	0.50	75	5	5	0.102	0.100	0.105	0.107	50	20	0.053	0.053	0.056	0.048
75	0.75	100	5	5	0.179	0.179	0.169	0.171	50	20	0.075	0.075	0.079	0.104
75	1.00	100	5	5	0.284	0.279	0.286	0.289	50	20	0.103	0.104	0.105	0.181*
100	0.25	50	5	5	0.058	0.055	0.059	0.058	50	20	0.036	0.039	0.040	0.022
100	0.50	75	5	5	0.107	0.104	0.103	0.102	50	20	0.053	0.053	0.053	0.054
100	0.75	75	5	5	0.184	0.184	0.183	0.183	50	20	0.075	0.080	0.081	0.109
100	1.00	100	5	5	0.286	0.288	0.286	0.291	50	20	0.103	0.110	0.110	0.183*

Table 3.4–continued.

Resource N=2000														
Parameter		Two-stage design							One-stage design					Improv.
ASP	ε	m	r	n_1	F2	Indep	Markov	Comb.	m	n	F1	Indep	Markov	F2-F1
10	0.25	50	5	5	0.028	0.027	0.028	0.029	50	10	0.028	0.026	0.025	0.000
10	0.50	50	5	5	0.036	0.033	0.036	0.039	50	10	0.037	0.037	0.040	-0.000
10	0.75	100	5	5	0.050	0.048	0.051	0.056	150	10	0.052	0.054	0.054	-0.002
10	1.00	225	5	5	0.082	0.084	0.081	0.090	200	10	0.090	0.095	0.087	-0.008
25	0.25	50	5	5	0.042	0.044	0.043	0.046	50	25	0.040	0.041	0.038	0.002
25	0.50	125	5	10	0.066	0.066	0.067	0.072	80	25	0.066	0.069	0.066	-0.000
25	0.75	175	5	10	0.136	0.130	0.133	0.143	80	25	0.117	0.122	0.111	0.019
25	1.00	175	5	10	0.246	0.241	0.245	0.259	80	25	0.189	0.198	0.192	0.057
50	0.25	50	5	15	0.053	0.048	0.045	0.050	50	40	0.048	0.046	0.048	0.005
50	0.50	100	14	10	0.113	0.115	0.111	0.117	50	40	0.079	0.075	0.079	0.034
50	0.75	150	5	10	0.226	0.225	0.226	0.232	60	33	0.122	0.121	0.126	0.104
50	1.00	150	5	10	0.385	0.392	0.372	0.381	80	25	0.189	0.187	0.191	0.197*
75	0.25	50	5	25	0.062	0.058	0.061	0.066	50	40	0.048	0.047	0.050	0.014
75	0.50	125	5	10	0.137	0.138	0.132	0.133	50	40	0.079	0.080	0.081	0.058
75	0.75	125	5	10	0.274	0.282	0.269	0.276	60	33	0.122	0.123	0.121	0.151*
75	1.00	125	5	10	0.438	0.435	0.432	0.437	80	25	0.189	0.185	0.186	0.249*
100	0.25	50	5	25	0.068	0.072	0.066	0.068	50	40	0.048	0.044	0.047	0.020
100	0.50	125	5	5	0.152	0.154	0.149	0.148	50	40	0.079	0.077	0.078	0.073
100	0.75	125	5	10	0.293	0.295	0.285	0.288	60	33	0.122	0.124	0.124	0.170*
100	1.00	125	5	10	0.456	0.445	0.445	0.449	80	25	0.189	0.192	0.186	0.267*

Table 3.4–continued.

Resource N=5000														
Parameter		Two-stage design							One-stage design					Improv.
ASP	ε	m	r	n_1	F2	Indep	Markov	Comb.	m	n	F1	Indep	Markov	F2-F1
10	0.25	50	5	5	0.028	0.029	0.028	0.031	50	10	0.028	0.029	0.029	0.000
10	0.50	50	5	5	0.036	0.036	0.035	0.039	50	10	0.037	0.040	0.037	-0.000
10	0.75	100	5	5	0.050	0.055	0.052	0.058	150	10	0.052	0.052	0.051	-0.002
10	1.00	225	5	5	0.082	0.084	0.080	0.089	240	10	0.089	0.085	0.088	-0.007
25	0.25	50	5	5	0.042	0.040	0.043	0.045	50	25	0.040	0.042	0.038	0.002
25	0.50	125	5	10	0.066	0.067	0.062	0.067	180	25	0.069	0.073	0.067	-0.004
25	0.75	225	27	5	0.141	0.141	0.132	0.144	200	25	0.159	0.160	0.145	-0.018
25	1.00	300	63	5	0.275	0.270	0.252	0.287	200	25	0.301	0.302	0.286	-0.026
50	0.25	50	5	15	0.053	0.043	0.049	0.056	50	50	0.052	0.052	0.056	0.000
50	0.50	200	43	5	0.136	0.141	0.132	0.138	100	50	0.129	0.135	0.130	0.007
50	0.75	250	29	10	0.325	0.333	0.306	0.321	100	50	0.269	0.275	0.260	0.055
50	1.00	250	29	10	0.572	0.570	0.532	0.560	100	50	0.457	0.464	0.446	0.115
75	0.25	75	40	5	0.062	0.050	0.052	0.065	60	75	0.063	0.064	0.058	-0.000
75	0.50	175	22	10	0.200	0.202	0.196	0.202	70	71	0.144	0.144	0.140	0.056
75	0.75	200	24	10	0.447	0.453	0.416	0.429	80	62	0.279	0.279	0.280	0.168*
75	1.00	200	24	15	0.701	0.698	0.659	0.687	100	50	0.457	0.459	0.445	0.245*
100	0.25	100	23	15	0.074	0.056	0.057	0.074	50	100	0.071	0.071	0.069	0.003
100	0.50	150	19	10	0.246	0.111	0.101	0.137	70	71	0.144	0.147	0.148	0.102
100	0.75	175	22	15	0.510	0.513	0.480	0.502	80	62	0.279	0.285	0.284	0.231*
100	1.00	125	17	20	0.757	0.745	0.744	0.764	100	50	0.457	0.465	0.459	0.300*

Table 3.4–continued.

Resource N=10,000														
Parameter		Two-stage design							One-stage design					Improv.
ASP	ε	m	r	n_1	F2	Indep	Markov	Comb.	m	n	F1	Indep	Markov	F2-F1
10	0.25	50	5	5	0.028	0.031	0.027	0.028	50	10	0.028	0.030	0.028	0.000
10	0.50	50	5	5	0.036	0.038	0.040	0.042	50	10	0.037	0.038	0.036	-0.000
10	0.75	100	5	5	0.050	0.054	0.051	0.057	150	10	0.052	0.057	0.054	-0.002
10	1.00	225	5	5	0.082	0.081	0.086	0.097	240	10	0.089	0.090	0.086	-0.007
25	0.25	50	5	5	0.042	0.042	0.043	0.043	50	25	0.040	0.038	0.041	0.002
25	0.50	125	5	10	0.066	0.066	0.063	0.068	180	25	0.069	0.072	0.071	-0.004
25	0.75	225	27	5	0.141	0.139	0.134	0.144	300	25	0.158	0.162	0.146	-0.017
25	1.00	300	63	5	0.275	0.274	0.244	0.270	350	25	0.315	0.316	0.270	-0.040
50	0.25	50	5	15	0.053	0.047	0.044	0.058	50	50	0.052	0.050	0.050	0.000
50	0.50	225	93	5	0.140	0.145	0.134	0.144	200	50	0.153	0.161	0.144	-0.013
50	0.75	250	101	5	0.345	0.351	0.319	0.344	200	50	0.368	0.383	0.353	-0.023
50	1.00	250	125	5	0.607	0.605	0.550	0.591	200	50	0.630	0.639	0.591	-0.023
75	0.25	100	68	5	0.062	0.045	0.047	0.058	130	75	0.064	0.067	0.060	-0.002
75	0.50	225	71	10	0.224	0.217	0.205	0.228	130	75	0.218	0.227	0.216	0.006
75	0.75	275	59	15	0.522	0.528	0.470	0.514	130	75	0.477	0.490	0.464	0.045
75	1.00	275	59	15	0.802	0.793	0.731	0.775	140	71	0.736	0.747	0.724	0.067
100	0.25	125	65	5	0.079	0.056	0.048	0.070	100	100	0.081	0.085	0.080	-0.002
100	0.50	225	49	15	0.303	0.238	0.231	0.285	100	100	0.255	0.267	0.259	0.047
100	0.75	225	49	15	0.645	0.657	0.594	0.622	100	100	0.526	0.541	0.522	0.119
100	1.00	250	29	25	0.886	0.876	0.825	0.859	100	100	0.780	0.785	0.784	0.106

3.4 Type I Error and Power of Claiming Linkage

Tables 3.3 and 3.4 give the optimal designs and the probabilities of finding the locus linked to the responsible gene when the gene exists, but they do not provide information on the probability of causing a false conclusion when there is no gene responsible for the disease. The usual requirement is that the LOD score in Stage II should be greater than a certain threshold, t, in order to claim linkage. Given $R = r$, we let the threshold be t, which is the 100(1-α) percentile of the unique maximum, T, of r Binomial(n_2, 0.25) random variables. The probability mass function of T is

$$\begin{aligned} & P(T = s) \\ = \; & P({}_2S_{I_0} = s | {}_2S_{I_0} \text{ is the unique maximum}) \\ = \; & \frac{P({}_2S_{I_0} = s, {}_2S_{I_0} \text{ is the unique maximum})}{P({}_2S_{I_0} \text{is the unique maximum})} \\ = \; & \frac{r f(s) F(s-1)^{r-1}}{\sum_{i=1}^{n_2} r f(s) F(s-1)^{r-1}}. \end{aligned}$$

The probability mass function of the marker group is

$$\begin{aligned} & P(T = s) \\ = \; & P({}_2S_{I_0} = {}_2S_{I_0+1} = s | {}_2S_{I_0} \text{ and } {}_2S_{I_0} \text{ are the unique maximum}) \\ = \; & \frac{P({}_2S_{I_0} = {}_2S_{I_0+1} = s, {}_2S_{I_0} \text{is the unique maximum})}{P({}_2S_{I_0} \text{ and } {}_2S_{I_0+1} \text{ are the unique maximum})} \\ = \; & \frac{(r-1) f(s)^2 F(s-1)^{r-2}}{\sum_{i=1}^{n_2} (r-1) f(s) F(s-1)^{r-2}}. \end{aligned}$$

Where $f(s)$ and $F(s)$ are probability mass function and cumulative distribution function of Binomial(n_2, 0.25). Table 3.5 and 3.6 gives the value of t of the unique maximum and the marker group. To use Table 3.5 and 3.6, first find your n_2 in the n_2 column, then find your R in R column, if your R is not in the table, then find the

largest value R that is smaller than your R with same n_2. The value in the t column is the 95% percentile. The range of n_2 is from 5 to 95, and for R is from 5 to 300.

A linkage can be claimed if and only if ${}_2S_{I_0} > t$, where I_0 is one of the markers that was chosen in Stage II. Clearly,

$$
\begin{aligned}
&P(\text{linkage claim is incorrect}) \\
&\leq P({}_2S_{I_0} > t \mid \text{no gene is responsible})P(\text{no gene is responsible}) \\
&\quad + P(I_0 \text{ is wrong and } {}_2S_{I_0} > t \mid \text{a gene is responsible}) \\
&\qquad P(\text{a gene is responsible}).
\end{aligned}
$$

The (prior) P(no gene is responsible) is usually unknown, but we can conclude that P(I_0 is wrong and ${}_2S_{I_0} > t \mid$ a gene is responsible) $\leq$ P(${}_2S_{I_0} > t \mid$ no gene is responsible). A proof is as follows. Let A denote the event { the gene is between marker i_0 and marker $i_0 + 1$}.

$$
\begin{aligned}
&P(I_0 \text{ is wrong and } {}_2S_{I_0} > t \mid \text{a gene is responsible}) && (3.42) \\
&= P({}_2S_{I_0} > {}_2S_j,\ \forall j \neq I_0,\ {}_2S_{I_0} > t \text{ and } I_0 \neq i_0 \text{ or } i_0 + 1 \mid A) && (3.43) \\
&= \sum_{k=t+1}^{n_2} P(k = {}_2S_{I_0} > {}_2S_j,\ \forall j \neq I_0, \text{ and } I_0 \neq i_0 \text{ or } i_0 + 1 \mid A) && (3.44) \\
&= \sum_{k=t+1}^{n_2} P(k > {}_2S_j,\ \forall j \neq I_0, i_0, \text{ or } i_0 + 1,\ k > {}_2S_{i_0},\ k > {}_2S_{i_0+1}, \\
&\qquad {}_2S_{I_0} = k \text{ and } I_0 \neq i_0 \text{ or } i_0 + 1 \mid A) && (3.45) \\
&= \sum_{k=t+1}^{n_2} \{P(k > {}_2S_j,\ \forall j \neq I_0, i_0, \text{ or } i_0 + 1,\ I_0 \neq i_0 \text{ or } i_0 + 1 \mid A) && (3.46) \\
&\qquad P(k > {}_2S_{i_0},\ k > {}_2S_{i_0+1} \mid A) && (3.47) \\
&\qquad P({}_2S_{I_0} = k,\ I_0 \neq i_0 \text{ or } i_0 + 1 \mid A)\}, && (3.48)
\end{aligned}
$$

Table 3.5. The 95% percentile of the unique maximum of R Binomial(n_2, 0.25).

n_2	R	t	n_2	R	t	n_2	R	t	n_2	R	t	n_2	R	t	n_2	R	t
5	5	5	28	44	18	45	20	21	59	145	28	72	255	33	84	286	37
5	24	5*	28	149	17	45	48	22	60	5	23	73	5	27	85	5	30
6	5	5	29	5	14	45	120	23	60	8	24	73	7	28	85	6	31
6	7	8	29	11	15	46	5	19	60	12	25	73	12	29	85	9	32
6	84	8*	29	29	16	46	7	20	60	22	26	73	22	30	85	15	33
7	5	6	29	92	17	46	15	21	60	48	27	73	42	31	85	28	34
7	22	7	30	5	14	46	34	22	60	108	28	73	88	32	85	50	35
8	5	8	30	8	15	46	85	23	60	284	29	73	193	33	85	99	36
8	9	7	30	20	16	46	233	24	61	5	24	74	6	27	85	207	37
8	89	8	30	80	17	47	5	19	61	10	25	74	6	28	86	5	31
9	5	7	30	202	18	47	8	20	61	18	26	74	10	29	86	8	32
9	24	8	31	5	14	47	12	21	61	37	27	74	18	30	86	13	33
9	225	9	31	8	15	47	28	22	61	82	28	74	34	31	86	22	34
10	5	7	31	15	16	47	62	23	61	194	29	74	69	32	86	41	35
10	11	8	31	40	17	47	162	24	62	5	24	74	148	33	86	79	36
10	70	9	31	128	18	48	5	20	62	8	25	75	5	28	86	162	37
11	5	7	32	5	15	48	10	21	62	15	26	75	9	29	87	5	31
11	8	8	32	11	16	48	20	22	62	29	27	75	15	30	87	7	32
11	29	9	32	28	17	48	48	23	62	63	28	75	28	31	87	11	33
11	218	10	32	81	18	48	115	24	62	144	29	75	55	32	87	18	34
12	5	8	32	273	19	49	5	20	63	5	24	75	115	33	87	33	35
12	14	9	33	5	15	49	8	21	63	7	25	75	258	34	87	84	36
12	79	10	33	8	16	49	18	22	63	12	26	76	5	28	87	128	37
13	5	8	33	20	17	49	34	23	63	23	27	76	7	29	87	271	38
13	8	9	33	54	18	49	83	24	63	49	28	76	13	30	88	5	31
13	36	10	33	171	19	49	219	25	63	108	29	76	23	31	88	8	32
13	231	11	34	5	15	50	5	20	63	259	30	76	44	32	88	14	33
14	5	9	34	8	16	50	8	21	64	5	24	76	90	33	88	18	34
14	19	10	34	14	17	50	12	22	64	6	25	76	198	34	88	28	35
14	95	11	34	36	18	50	28	23	64	10	26	77	5	28	88	52	36
15	5	9	34	111	19	50	81	24	64	19	27	77	8	29	88	102	37
15	11	10	35	5	18	50	155	25	64	38	28	77	11	30	88	212	38
15	46	11	35	11	19	51	5	21	64	83	29	77	19	31	89	5	32
15	285	12	35	27	18	51	10	22	64	192	30	77	35	32	89	8	33
16	5	9	35	74	19	51	20	23	65	5	25	77	71	33	89	13	34
16	7	10	35	232	20	51	46	24	65	8	26	77	151	34	89	23	35
16	25	11	36	5	16	51	112	25	65	15	27	78	5	28	89	43	36
16	119	12	36	8	17	51	296	28	65	30	28	78	6	29	89	82	37
17	5	10	36	19	18	52	5	21	65	64	29	78	9	30	89	167	38
17	15	11	36	51	19	52	8	22	65	144	30	78	16	31	90	5	32
17	60	12	36	150	20	52	16	23	66	5	25	78	29	32	90	7	33
18	5	10	37	5	16	52	35	24	66	7	26	78	57	33	90	11	34
18	9	11	37	7	17	52	82	25	66	13	27	78	118	34	90	20	35
18	33	12	37	14	18	52	209	26	66	24	28	78	258	35	90	35	36
18	152	13	37	36	19	53	5	21	66	50	29	79	5	29	90	67	37
19	5	10	37	100	20	53	7	22	66	109	30	79	6	30	90	133	38
19	6	11	38	5	17	53	13	23	66	258	31	79	13	31	90	277	39
19	20	12	38	11	18	53	24	24	67	5	25	79	24	32	91	5	32
19	79	13	38	26	19	53	61	25	67	6	26	79	46	33	91	8	33
20	5	11	38	69	20	53	150	26	67	11	27	79	93	34	91	10	34
20	13	12	38	205	21	54	5	21	67	20	28	79	199	35	91	17	35
20	45	13	39	5	17	54	8	22	67	39	29	80	5	29	91	29	36
20	198	14	39	9	18	54	10	23	67	84	30	80	7	30	91	54	37
21	5	11	39	19	19	54	21	24	67	192	31	80	11	31	91	106	38
21	9	12	39	48	20	54	46	25	68	5	28	80	20	32	91	217	39
21	27	13	39	136	21	54	110	26	68	9	27	80	37	33	92	5	32
21	105	14	40	5	18	54	282	27	68	16	28	80	74	34	92	6	33
22	5	11	40	7	18	55	5	22	68	31	29	80	154	35	92	9	34
22	6	12	40	15	19	55	9	23	68	65	30	81	5	29	92	14	35
22	17	13	40	35	20	55	17	24	68	145	31	81	6	30	92	24	36
22	80	14	40	93	21	55	35	25	69	5	26	81	10	31	92	45	37
22	256	15	40	278	22	55	61	26	69	8	27	81	17	32	92	88	38
23	5	12	41	5	17	55	202	27	69	13	28	81	30	33	92	172	39
23	12	13	41	8	18	56	5	22	69	25	29	81	59	34	93	5	33
23	37	14	41	11	19	56	7	23	69	51	30	81	121	35	93	8	34
23	140	15	41	25	20	56	13	24	69	111	31	81	262	36	93	12	35
24	5	12	41	65	21	56	28	25	69	255	32	82	5	30	93	21	36
24	8	13	41	185	22	56	61	26	70	5	26	82	6	31	93	37	37
24	24	14	42	5	18	56	147	27	70	7	27	82	14	32	93	70	38
24	82	15	42	9	19	57	5	22	70	11	28	82	25	33	93	137	39
25	5	12	42	19	20	57	6	23	70	21	29	82	48	34	93	284	40
25	8	13	42	47	21	57	11	24	70	41	30	82	98	35	94	5	33
25	18	14	42	127	22	57	22	25	70	86	31	82	202	36	94	7	34
25	50	15	43	5	18	57	47	26	70	192	32	83	5	30	94	11	35
25	188	16	43	7	19	57	108	27	71	5	26	83	7	31	94	16	36
26	5	13	43	15	20	57	272	28	71	6	27	83	12	32	94	31	37
26	11	14	43	34	21	58	5	23	71	9	26	83	21	33	94	57	38
26	32	15	43	89	22	58	9	24	71	17	28	83	39	34	94	110	39
26	110	16	43	252	23	58	17	25	71	33	30	83	76	35	94	224	40
27	5	13	44	5	18	58	36	26	71	67	31	83	158	36	95	5	33
27	8	14	44	6	19	58	81	27	71	146	32	84	5	30	95	6	34
27	22	15	44	12	20	58	197	28	72	5	27	84	6	31	95	9	35
27	88	16	44	28	21	59	5	23	72	8	28	84	10	32	95	15	36
27	251	17	44	63	22	59	8	24	72	14	29	84	18	33	95	26	37
28	5	13	44	172	23	59	14	25	72	26	30	84	32	34	95	47	38
28	6	14	45	5	19	59	28	26	72	53	31	84	61	35	95	89	39
28	15	15	45	9	20	59	62	27	72	113	32	84	124	36	95	178	40

Table 3.6. The 95% percentile of the unique maximum marker group of R Binomial(n_2, 0.25).

n_2	R	t	n_2	R	t	n_2	R	t	n_2	R	t	n_2	R	t	n_2	R	t
5	5	4	28	64	15	45	8	18	59	108	26	72	95	30	84	113	34
5	12	5	28	203	16	45	16	19	59	248	27	72	201	31	84	227	35
5	173	5*	29	5	12	45	53	20	60	5	21	73	5	25	85	5	28
6	5	4	29	8	13	45	78	21	60	7	22	73	8	26	85	8	29
6	6	5	29	17	13	45	193	22	60	11	23	73	12	27	85	10	30
6	38	6	29	44	15	46	5	17	60	20	24	73	21	28	85	18	31
7	5	5	29	131	16	46	7	16	60	40	25	73	39	29	85	27	32
7	16	6	30	5	12	46	13	19	60	83	26	73	76	30	85	49	33
7	132	7	30	8	13	46	26	20	60	187	27	73	158	31	85	92	34
8	5	5	30	13	15	46	57	21	61	5	21	74	5	25	85	181	35
8	8	6	30	31	15	46	140	22	61	8	22	74	7	28	86	5	28
8	43	7	30	87	16	47	5	17	61	10	23	74	11	27	86	8	29
9	5	6	30	278	17	47	6	18	61	17	24	74	18	28	86	15	30
9	21	7	31	5	13	47	10	19	61	32	25	74	32	29	86	14	31
9	130	8	31	10	14	47	21	20	61	85	26	74	62	30	86	23	32
10	5	6	31	23	15	47	44	21	61	143	27	74	125	31	86	41	33
10	12	7	31	80	16	47	103	22	62	5	22	74	285	32	86	75	34
10	54	8	31	179	17	47	262	23	62	8	23	75	5	25	86	148	35
11	5	6	32	5	13	48	5	18	62	14	24	75	6	26	86	296	38
11	7	7	32	8	14	48	9	19	62	26	25	75	9	27	87	5	29
11	28	8	32	17	15	48	17	20	62	52	26	75	15	28	87	8	30
11	151	9	32	43	16	48	34	21	62	110	27	75	27	29	87	12	31
12	5	7	32	120	17	48	77	22	62	249	28	75	50	30	87	20	32
12	16	8	33	5	13	48	189	23	63	5	22	75	99	31	87	34	33
12	70	9	33	8	14	49	5	18	63	7	23	75	207	32	87	62	34
13	5	7	33	13	15	49	7	19	63	12	24	76	5	26	87	118	35
13	10	8	33	31	16	49	13	20	63	22	25	76	8	27	87	235	36
13	37	9	33	82	17	49	27	21	63	42	26	76	13	28	88	5	29
13	167	10	33	246	18	49	56	22	63	86	27	76	23	29	88	7	30
14	5	7	34	5	14	49	136	23	63	189	28	76	41	30	88	10	31
14	7	8	34	10	15	50	5	16	64	5	22	76	80	31	88	17	32
14	22	9	34	23	18	50	6	19	64	6	23	76	163	32	88	29	33
14	93	10	34	50	17	50	11	20	64	10	24	77	5	26	88	51	34
15	5	6	34	164	18	50	21	21	64	18	25	77	7	27	88	98	35
15	14	9	34	5	14	50	42	22	64	34	26	77	11	28	88	166	36
15	51	10	35	8	15	50	103	23	64	68	27	77	19	29	89	5	29
15	240	11	35	18	16	50	255	24	64	146	28	77	34	30	89	6	30
16	5	6	35	42	17	51	5	19	65	5	23	77	65	31	89	9	31
16	9	9	35	113	18	51	9	20	65	9	24	77	129	32	89	15	32
16	30	10	36	5	14	51	17	21	65	15	25	77	272	33	89	24	33
16	124	11	36	7	15	51	35	22	65	28	26	78	5	26	89	43	34
17	5	8	36	14	16	51	76	23	65	54	27	78	6	27	89	79	35
17	7	9	36	31	18	51	186	24	65	113	28	78	10	28	89	152	36
17	19	10	36	79	18	52	5	19	65	252	29	78	16	29	90	5	29
17	69	11	36	225	19	52	8	20	66	5	23	78	29	30	90	6	30
18	5	9	37	5	14	52	14	21	66	8	24	78	53	31	90	8	31
18	13	10	37	6	15	52	28	22	66	13	25	78	103	32	90	13	32
18	41	11	37	11	16	52	60	23	66	23	26	78	213	33	90	21	33
18	166	12	37	24	17	52	138	24	66	44	27	79	5	26	90	36	34
19	5	9	37	57	18	53	5	19	66	89	28	79	6	27	90	65	35
19	9	10	37	154	19	53	7	20	66	193	29	79	9	28	90	123	36
19	26	11	38	5	15	53	12	21	67	5	23	79	14	29	90	244	37
19	94	12	38	9	16	53	22	22	67	7	24	79	24	30	91	5	30
20	5	9	38	18	17	53	47	23	67	11	25	79	44	31	91	7	31
20	11	10	38	42	18	53	104	24	67	19	26	79	83	32	91	11	32
20	18	11	38	106	18	53	251	25	67	35	27	79	166	33	91	18	33
20	57	12	39	5	15	54	5	19	67	70	28	80	5	27	91	31	34
20	223	13	39	7	16	54	9	20	67	149	29	80	8	28	91	54	35
21	5	10	39	14	17	54	10	21	68	5	23	80	12	29	91	101	36
21	13	11	39	32	18	54	18	22	68	6	24	80	20	30	91	196	37
21	36	12	39	77	18	54	37	23	68	9	25	80	36	31	92	5	30
21	129	13	39	210	20	54	79	24	68	16	26	80	66	32	92	7	31
22	5	10	40	5	16	54	185	25	68	29	27	80	134	33	92	12	32
22	9	11	40	6	16	55	5	20	68	56	28	80	279	34	92	16	33
22	24	12	40	12	17	55	8	21	68	116	29	81	5	27	92	26	34
22	78	13	40	24	18	55	15	22	68	255	30	81	7	28	92	45	35
23	5	10	40	57	19	55	29	23	69	5	24	81	11	29	92	83	36
23	7	11	40	147	20	55	54	24	69	8	25	81	17	30	92	159	37
23	17	12	41	5	16	55	139	25	69	14	26	81	30	31	93	5	30
23	50	13	41	9	17	56	5	20	69	24	27	81	56	32	93	8	31
23	176	14	41	19	18	56	7	21	69	46	28	81	106	33	93	14	32
24	5	10	41	42	19	56	13	22	69	92	29	81	220	34	93	14	33
24	6	11	41	105	20	56	24	23	69	197	30	82	5	27	93	22	34
24	13	12	41	286	21	56	48	24	70	5	24	82	6	28	93	36	35
24	34	13	42	5	16	56	105	25	70	7	25	82	9	29	93	69	36
24	106	14	42	7	17	56	246	26	70	12	26	82	15	30	93	129	37
25	5	11	42	15	18	57	5	20	70	20	27	82	26	31	93	254	38
25	9	12	42	32	19	57	6	21	70	37	28	82	46	32	94	5	31
25	23	13	42	77	20	57	11	22	70	73	29	82	67	33	94	6	32
25	69	14	42	200	21	57	19	23	70	153	30	82	174	34	94	12	33
25	240	15	43	5	17	57	34	24	71	5	24	83	5	28	94	19	34
26	5	11	43	7	17	57	81	25	71	6	25	83	8	29	94	32	35
26	7	12	43	17	18	57	186	26	71	10	26	83	13	30	94	57	36
26	17	13	43	25	19	58	5	21	71	17	27	83	22	31	94	106	37
26	46	14	43	57	20	58	9	22	71	31	28	83	36	32	94	204	38
26	148	15	43	142	21	58	18	23	71	59	29	83	71	33	95	5	31
27	5	11	44	5	16	58	31	24	71	120	30	83	140	34	95	7	32
27	6	12	44	6	17	58	63	25	71	260	31	83	287	35	95	11	33
27	13	13	44	10	18	58	140	26	72	5	24	84	5	28	95	17	34
27	32	14	44	20	19	59	5	21	72	6	25	84	7	29	95	28	35
27	95	15	44	43	20	59	8	22	72	9	26	84	11	30	95	46	36
28	5	12	44	103	21	59	13	23	72	14	27	84	18	31	95	67	37
28	10	13	44	272	22	59	25	24	72	25	28	84	32	32	95	166	38
28	23	14	45	5	17	59	50	25	72	48	29	84	59	33			

and

$$P({}_2S_{I_0} > t \mid \text{no gene is responsible}) \tag{3.49}$$

$$= \sum_{k=t+1}^{n_2} \{P(k > {}_2S_j,\ j \neq I_0, i_0 \text{ or } i_0+1 \text{ and } I_0 \neq i_0 \text{ or } i_0+1 \mid \text{no gene}) \tag{3.50}$$

$$P(k > {}_2S_{i_0},\ k > {}_2S_{i_0+1} \mid \text{no gene}) \tag{3.51}$$

$$P({}_2S_{I_0} = k \text{ and } I_0 \neq i_0 \text{ or } i_0+1 \mid \text{no gene})\}. \tag{3.52}$$

Eq. (3.47) is less than Eq. 3.51 and the other corresponding terms are equal, therefore

$$P(I_0 \text{ is wrong and } {}_2S_{I_0} > t \mid \text{a gene is responsible})$$
$$\leq P({}_2S_{I_0} > t \mid \text{no gene is responsible})$$

and thus,

$$P(\text{linkage claim is incorrect})$$
$$\leq P({}_2S_{I_0} > t \mid \text{no gene is responsible})$$
$$= \sum_{\substack{\text{all possible results} \\ \text{of the stage I}}} \left\{ \left[\sum_{i=0}^{r} P({}_2S_i > t_{\alpha,n_2} \mid I_0 = i) P(I_0 = i) \right] P(\text{the result of stage I}) \right\}$$
$$= \alpha \sum_{i=0}^{m} P(I_0 = i) \leq \alpha.$$

3.5 Discussion

The Monte Carlo results indicated that the relative errors between the probabilities calculated from formulas and Markov chain simulations were under 7% in the dominant cases and under 15% in the recessive cases. Consequently, the approximation using the independence assumption for dependent marker loci was acceptable.

The simulation studies also showed that, in the dominant cases, combining Stage I data with Stage II data did not have any significant advantage. The probability

of allocating the correct marker increased less than by 3%. For the recessive cases, there was some advantage when the ASPs were few. The probability of allocating the correct marker could increase by as much as 10%. However, it is very diffcult to combine data in the theoretical derivation.

The powers to find the gene depends on the exact gene location between markers. Since Table 3.3 and 3.4 were constructed under the least favorable configuration, the actual power should be higher.

The two-stage approach indeed boosted the probability of finding the correct gene location under resource constraints. In many cases this probability can increase up to 20% to 30%. In searching for recessive disease genes, there were several instances when the improvement exceeded 35% (Table 3.3). However, if there are enough resources, such that almost all available markers can be typed on all ASPs, then the one-stage approach may have higher power. (see e.g., N=5000 and APS=10 or N =10,000 and ASP=10, 25 in Table 3.3.) However, since the power loss is so small, we may always choose the optimal two-stage design.

As shown in Tables 3.3 and 3.4, it requires much more resources to locate a dominant than a recessive disease gene. For example, in a two-stage design with 25 ASPs and N=1000 we can locate a recessive disease gene with a 0.88 probability when $\varepsilon = 1$. But for a dominant disease gene, more than 100 ASPs and/or more than N=10,000 are needed to achieve the same probability.

Another point worth noting is that phenocopy can severely reduce the probability of finding correct gene location. For example, to locate a recessive gene, with resource N=5000, 25 ASPs, and $\varepsilon = 1$, the chance of finding it is 99.9%. However, when $\varepsilon = 0.5$, even if we double the resource to N=10,000 and 50 ASPs, the chance of finding the correct locus is only 90%. Althought both of them have about 25 ASPs whose disease is caused by gene, those extra phenocopy ASPs reduce the probability considerably.

CHAPTER 4
TWO-STAGE GENOME SEARCH FOR COMPLEX DISEASE

In the previous chapter, we have focused on finding a single disease gene. However, genetic diseases are not always single-gene diseases. Many of them are complex diseases, i.e., diseases caused by several genes. For example, insulin-dependent diabetes mellitus (IDDM) is influenced by a number of susceptibility genes and environmental factors (Luo et al., 1995). A disease phenotype controlled by genes at several different loci is considered to have "nonallelic heterogeneity" and "when this disease is relatively rare and mutation rates are low, individuals within a family are generally homogeneous. Locus heterogeneity then leads to the situation that the recombination fraction between disease phenotype and marker will be different in different families" (Ott, 1991, p. 199). This chapter discusses the probability to find two unlinked recessive genes that may cause the same disease in a two-stage search under the following genetic model.

4.1 Genetic Model and Assumptions

Throughout this chapter, we assumed the following:

- No epistasis, i.e. no interaction between genes.
- When an individual carries the two disease genes, the penetrance is additive, i.e. the probability of being affected for an individual who has two recessive genes at two loci is twice as high as for an individual who has two recessive genes at only one locus.

- For an individual who has no disease gene, the probability of being affected is k_0 times the probability of being affected for an individual who has two recessive genes at one locus.
- Receiving one gene has no effect on receiving other genes.

Other assumptions, necessary for simplifying analytic study are similar to those in the previous chapter, are:

- There are two alleles at each gene locus, denote disease gene d and normal gene D. The population frequency of disease gene is p for both loci.
- Genes are in the middle of two adjacent markers.
- In this study, only three maps, 5 cM (centimorgan), 10 cM, and 20 cM, are available, and every marker are highly polymorphic. Marker's positions on the each chromosome are at 0 cM, 5 cM, 10 cM,..., for 5 cM map, 0 cM, 10 cM,..., for 10cM map, and 0 cM, 20 cM, 40 cM,..., for 20 cM map.
- The cost of typing alleles is a constant.

4.2 Two-stage Genome Search

For simplicity, we use a slightly different two-stage approach for searching complex disease genes. In the first stage, we choose a threshold for statistics instead of choosing a number of loci. In the second stage, we choose the loci on different chromosomes where the statistics have the highest overall value. Details are as follows:

Following chapter 3, suppose there are n ASPs and there are enough resources to type N markers for ASPs. Again, there are three numbers that must be determined in a two-stage design: n_1 and m, the number of ASPs and the number of markers to be used in Stage I, and k, the threshold of markers to be studied in Stage II. Define

$$_1S_i = \sum_{j=1}^{n_1} X_{ij}. \tag{4.1}$$

in the same way as in chapter 3 where i is the index of loci, j is the index of ASP, X_{ij} is defined by the assumption 6 in §3.1. If in Stage I, marker i has a ${}_1S_i$ value higher than the threshold, k, we will study this marker again in the stage II. Let the number of markers that passed Stage I be R.

In Stage II, R markers on N_2 ASPs are to be typed, where N_2 is the largest number subject to the resource constraint and R. Since R is a random variable, N_2 is also a random variable. Thus, N_2 is the largest x, such that $mn_1 + Rx \leq N$. We define

$$ {}_2S_i = \sum_{j=n_1+1}^{n_1+N_2} X_{ij}, \tag{4.2} $$

for stage II. Then, markers that meet the following criteria will be declared having a gene nearby:

- A marker with the uniquely highest ${}_2S_i$ is claimed to be the marker nearest to the disease gene.
- A marker group (see page 42), has the uniquely highest score, is claimed to have the gene lie between them,
- If two markers or marker groups have the same highest score but on different chromosomes, then declare each having a gene nearby in the corresponding way in the above.

If none of the above applies, then gene location is considered undetermined.

Once the location(s) has(have) been chosen, same as in the chapter 3, the next step is to check whether we can claim linkage. Let t be the 100(1 - α) percentile of the maximum of r binomial (n_2, 0.25) random variables. If ${}_2S_i$ of the chosen location(s) is(are) greater than t, then we claim there is linkage at that location(s). Since we imposed some restriction on declaring a maker having a gene nearby, the actual type I error will be less than α. The 95% percentile for the unique maximum and marker group were given in Table 3.5 and 3.6, respectively.

4.3 Probability of Allocating the Correct Marker for a Complex Disease

In this section we discuss an analytical approach and some simulation results. §4.3.1 is a preparation for the discussion.

4.3.1 Possible Parental Genotypes and Trait IBD distribution

For a disease gene locus, say locus i, let d and D denote the disease and the normal gene allele respectively. Let also the population frequency of d be p, and PG denote the parental genotype. Let

$$E_i = (e_{i1}, e_{i2}) =$$

$$= \begin{cases} (1,1), & \text{if both sibs received two recessive genes at locus } i \\ (1,0), & \text{if sib 1 received two recessive genes and sib 2 did not at locus } i \\ (0,1), & \text{if sib 2 received two recessive genes and sib 1 did not at locus } i \\ (0,0), & \text{if none of sibs received two recessive genes at locus } i. \end{cases}$$

Then the distribution of trait IBD, I_t, conditional on parental genotype and E_i is given in Table 4.1. The numbers in the table is very easy to verify, for example, for parental genotype 2, if the order of the chromosomes are specified, say 1, 2, 3, and 4 where 1 and 2 belong to father, and 3 and 4 belong to mother, then the probability of receiving Dd dd, where D is on the chromosome 4, is $p^3(1-p)$. If we permute the order of chromosomes we will get 4 different permutation, therefore, the probability of parents having genotype Dd and dd is $4p^3(1-p)$ in a random mating population. When parental genotype is Dd and dd, possible sib genotype are Dd and dd with equal frequency, therefore, P(E_i=(1,1)|PG=2)= P(E_i=(1,0)|PG=2)= P(E_i=(0,1)|PG=2)=P(E_i=(0,0)|PG=2)= $\frac{1}{4}$. In order to illustrate the conditional distribution of trait IBD score, subscripts for parental genotype are denoted as d_1d_2 and Dd_3. When parental genotype is dd and Dd, and $E_i = (0,0)$, sib's genotype

can only be Dd_1 or Dd_2 with equal chance, therefore, $\mathrm{P}(I_t{=}2|\mathrm{PG}{=}2, E_i{=}(0,0)) = \mathrm{P}(I_t{=}1|\mathrm{PG}{=}2, E_i{=}(0,0)){=}\frac{1}{2}$.

4.3.2 Analytic Approach

Assume there are 2 unlinked genes, namely G_1 and G_2. Without loss generality, let X_{1j} and X_{2j} be the statistics of the markers next to G_1, and X_{3j} and X_{4j} be the statistics of the markers next to G_2 for the ASP j; ${}_lS_1$, ${}_lS_2$, ${}_lS_3$, and ${}_lS_4$ be the sum of Xs respectively in Stage l, as defined at the beginning of this section. Also, let I_1 and I_2 be the IBD score of G_1 and G_2 for the affected sibs pair. Let the penetrance of carrying two recessive gene at one locus is λ, two loci is 2λ, and none is $k_0\lambda$, where k_0 is the relative risk of being affected for an individual carrying no gene to an individual carrying two recessive genes at one locus. The λ is unknown but is assumed to be small. It will be cancel out in the formula. Then

$$
\begin{aligned}
& P(\text{a gene or both genes are found}) \\
= \; & P(G_1 \text{ is found and } G_2 \text{ is not}) \\
& +P(G_2 \text{ is found and } G_1 \text{ is not}) \\
& +P(G_1 \text{ and } G_2 \text{ are found}) \\
= \; & \{P(G_1 \text{ passes Stage I, } G_2 \text{ does not pass Stage I, } G_1 \text{ is found}) && (4.3) \\
& +P(G_1 \text{ and } G_2 \text{ pass Stage I, } G_1 \text{ is found})\} && (4.4) \\
& +\{P(G_2 \text{ passes Stage I, } G_1 \text{ does not pass I,} G_2 \text{ is found}) && (4.5) \\
& +P(G_1 \text{ and } G_2 \text{ pass Stage I, } G_2 \text{ is found})\} && (4.6) \\
& +P(G_1 \text{ and } G_2 \text{ pass Stage I, } G_1 \text{ and } G_2 \text{ are found}) && (4.7)
\end{aligned}
$$

For a given threshold k, a given event $\{G_1$ is found but G_2 is not$\}$, the possible relationship between k and ${}_1S_1$, ${}_1S_2$, ${}_1S_3$, and ${}_1S_4$ in Stage I, and corresponding relationship of ${}_2S_1$, ${}_2S_2$, ${}_2S_3$, and ${}_2S_4$ in Stage II are given in Table 4.2. For example for event 6, given G_1 is found but G_2 is not, in Stage I ${}_1S_1 > k$, ${}_1S_2 > k$, ${}_1S_3 \leq k$

Table 4.1. Trait IBD distribution conditional on parental genotype and E vector.

PG	P(PG)	P(E=(1,1) | PG)	P(E=(1,0) | PG)	P(E=(0,1) | PG)	P(E=(0,0) | PG)
1. *dd dd*	p^4	1 P(I_t=2 | PG, E)=1	0	0	0
2. *Dd dd*	$4p^3(1-p)$	$\frac{1}{4}$ P(I_t=2 | PG, E)=1	$\frac{1}{4}$ P(I_t=1 | PG, E)=$\frac{1}{2}$ P(I_t=0 | PG, E)=$\frac{1}{2}$	$\frac{1}{4}$ P(I_t=1 | PG, E)=$\frac{1}{2}$ P(I_t=0 | PG, E)=$\frac{1}{2}$	$\frac{1}{4}$ P(I_t=2 | PG, E)=$\frac{1}{2}$ P(I_t=1 | PG, E)=$\frac{1}{2}$
3. *Dd Dd*	$4p^2(1-p)^2$	$\frac{1}{16}$ P(I_t=2 | PG, E)=1	$\frac{3}{16}$ P(I_t=1 | PG, E)=$\frac{2}{3}$ P(I_t=0 | PG, E)=$\frac{1}{3}$	$\frac{3}{16}$ P(I_t=1 | PG, E)=$\frac{2}{3}$ P(I_t=0 | PG, E)=$\frac{1}{3}$	$\frac{9}{16}$ P(I_t=2 | PG, E)=$\frac{3}{9}$ P(I_t=1 | PG, E)=$\frac{4}{9}$ P(I_t=0 | PG, E)=$\frac{2}{9}$
4.*DD dd*	$2p^2(1-p)^2$	0	0	0	1 P(I_t=2 | PG, E)=$\frac{1}{4}$ P(I_t=1 | PG, E)=$\frac{1}{2}$ P(I_t=0 | PG, E)=$\frac{1}{4}$
5. *DD Dd*	$4p(1-p)^3$	0	0	0	1 P(I_t=2 | PG, E)=$\frac{1}{4}$ P(I_t=1 | PG, E)=$\frac{1}{2}$ P(I_t=0 | PG, E)=$\frac{1}{4}$
6. *DD DD*	$(1-p)^4$	0	0	0	1 P(I_t=2 | PG, E)=$\frac{1}{4}$ P(I_t=1 | PG, E)=$\frac{1}{2}$ P(I_t=0 | PG, E)=$\frac{1}{4}$

Table 4.2. Exclusive events for the case "G_1 is found but G_2 is not."

	Stage I				Stage II
event	${}_1S_1$	${}_1S_2$	${}_1S_3$	${}_1S_4$	the largest statistics and relationship
1	$>$	$>$	$\leq$	$\leq$	$\max({}_2S_1, {}_2S_2)$
2	$>$	$\leq$	$\leq$	$\leq$	${}_2S_1$
3	$\leq$	$>$	$\leq$	$\leq$	${}_2S_2$
4	$>$	$>$	$>$	$>$	$\max({}_2S_1, {}_2S_2)$, and $> \max({}_2S_3, {}_2S_4)$
5	$>$	$>$	$>$	$\leq$	$\max({}_2S_1, {}_2S_2)$, and $> {}_2S_3$
6	$>$	$>$	$\leq$	$>$	$\max({}_2S_1, {}_2S_2)$, and $> {}_2S_4$
7	$>$	$\leq$	$>$	$>$	${}_2S_1$, and $> \max({}_2S_3, {}_2S_4)$
8	$\leq$	$>$	$>$	$>$	${}_2S_2$, and $> \max({}_2S_3, {}_2S_4)$
9	$>$	$\leq$	$>$	$\leq$	${}_2S_1$ and $> {}_2S_3$
10	$>$	$\leq$	$\leq$	$>$	${}_2S_1$, and $> {}_2S_4$
11	$\leq$	$>$	$>$	$\leq$	${}_2S_2$, and $> {}_2S_3$
12	$\leq$	$>$	$\leq$	$>$	${}_2S_2$, and $> {}_2S_4$

and ${}_1S_4 > k$, then in the Stage II, the maximum of ${}_2S_1$ and ${}_2S_2$ must have the largest value and also larger than ${}_2S_4$.

For a given threshold k, and event $\{G_2$ is found but G_1 is not$\}$ the possible relationship between k and ${}_1S_1$, ${}_1S_2$, ${}_1S_3$, and ${}_1S_4$ in Stage I, and corresponding relationship in Stage II are given in Table 4.3.

For a given threshold k, and event $\{G_1$ and G_2 are found$\}$ the possible relationship between k and ${}_1S_1$, ${}_1S_2$, ${}_1S_3$, and ${}_1S_4$ in Stage I, and corresponding relationship in Stage II are given in Table 4.4.

For a given threshold k, the sum of probabilities (4.3) through (4.7) are

$$\sum_{l=0}^{m-4} \sum_{i=1}^{33} \sum_{h=1}^{N_2} P(\text{event i, and } l \ {}_1S_j \text{ pass Stage I, the largest statistics } = h \text{ in Stage II }),$$

Table 4.3. Exclusive events for the case "G_2 is found but G_1 is no."

	Stage I				Stage II
event	${}_1S_1$	${}_1S_2$	${}_1S_3$	${}_1S_4$	the largest statistics and relationship
13	$\leq$	$\leq$	$>$	$>$	$\max({}_2S_3, {}_2S_4)$
14	$\leq$	$\leq$	$>$	$\leq$	${}_2S_3$
15	$\leq$	$\leq$	$\leq$	$>$	${}_2S_4$
16	$>$	$>$	$>$	$>$	$\max({}_2S_3, {}_2S_4)$, and $> \max({}_2S_1, {}_2S_2)$
17	$>$	$\leq$	$>$	$>$	$\max({}_2S_3, {}_2S_4)$, and $> {}_2S_1$
18	$\leq$	$>$	$>$	$>$	$\max({}_2S_3, {}_2S_4)$, and $> {}_2S_2$
19	$>$	$>$	$>$	$\leq$	${}_2S_3$, and $> \max({}_2S_1, {}_2S_2)$
20	$>$	$>$	$\leq$	$>$	${}_2S_4$, and $> \max({}_2S_1, {}_2S_2)$
21	$>$	$\leq$	$>$	$\leq$	${}_2S_3$, and $> {}_2S_1$
22	$>$	$\leq$	$\leq$	$>$	${}_2S_4$, and $> {}_2S_1$
23	$\leq$	$>$	$>$	$\leq$	${}_2S_3$, and $> {}_2S_2$
24	$\leq$	$>$	$\leq$	$>$	${}_2S_4$, and $> {}_2S_2$

Table 4.4. Exclusive events for the case "G_1 and G_2 are found."

	Stage I				Stage II
event	${}_1S_1$	${}_1S_2$	${}_1S_3$	${}_1S_4$	the largest statistics and relationship
25	$>$	$>$	$>$	$>$	$\max({}_2S_3, {}_2S_4) = \max({}_2S_1, {}_2S_2)$
26	$>$	$\leq$	$>$	$>$	$\max({}_2S_3, {}_2S_4) = {}_2S_1$
27	$\leq$	$>$	$>$	$>$	$\max({}_2S_3, {}_2S_4) = {}_2S_2$
28	$>$	$>$	$>$	$\leq$	$\max({}_2S_1, {}_2S_2) = {}_2S_3$
29	$>$	$>$	$\leq$	$>$	$\max({}_2S_1, {}_2S_2) = {}_2S_4$
30	$>$	$\leq$	$>$	$\leq$	${}_2S_3 = {}_2S_1$
31	$>$	$\leq$	$\leq$	$>$	${}_2S_4 = {}_2S_1$
32	$\leq$	$>$	$>$	$\leq$	${}_2S_3 = {}_2S_2$
33	$\leq$	$>$	$\leq$	$>$	${}_2S_4 = {}_2S_2$

where $N_2 = \left[\frac{N-mn_1}{R}\right]_{Int}$.

For example for event 25;

$$
\begin{aligned}
& P(\text{event 25, and } l\ {}_1S_j\text{s pass Stage I}) \\
= \; & P({}_1S_1 > k,\ {}_1S_2 > k,\ {}_1S_3 > k,\ {}_1S_4 > k,\ l\ {}_1S_j > k \\
& \max({}_2S_3,\ {}_2S_4) = \max({}_2S_1,\ {}_2S_2) = h, \text{ and all other } l\ {}_2S_j\text{s} < h) \\
& \text{(with assumption of independence)} \\
= \; & P({}_1S_1 > k,\ {}_1S_2 > k,\ {}_1S_3 > k,\ {}_1S_4 > k) P(l\ {}_1S_j \text{ pass Stage I}) \\
& P(\max({}_2S_3,\ {}_2S_4) = \max({}_2S_1,\ {}_2S_2) = h) P(\text{all other } l\ {}_2S_j\text{s} < h),
\end{aligned}
$$

where $j \neq 1, 2, 3, 4$. The distribution of ${}_2S_j$ is Binomial(N_2, 0.25). Therefore, if we want to know probabilities (4.3) through (4.7), we need to know the joint distribution of ${}_1S_1$, ${}_1S_2$, ${}_1S_3$ and ${}_1S_4$ conditional on both sibs are affected. In order to calculate this joint distribution, we need to know the joint distribution of X_{1j} and X_{2j} and X_{3j} and X_{4j} conditional on both sibs are affected, which is,

$$p_{xyuv} \tag{4.8}$$

$$= P(X_{1j} = x, X_{2j} = y, X_{3j} = u, X_{4j} = v \mid \text{both affected}) \tag{4.9}$$

$$= \sum_{i_1=0}^{2} \sum_{i_2=0}^{2} P(X_{1j} = x, X_{2j} = y, X_{3j} = u, X_{4j} = v \ I_1 = i_1,\ I_2 = i_2 \mid \text{both affected})$$

$$
\begin{aligned}
= \; & \sum_{i_1=0}^{2} \sum_{i_2=0}^{2} \{P(X_{1j} = x, X_{2j} = y, X_{3j} = u, X_{4j} = v \mid I_1 = i_1,\ I_2 = i_2, \text{ both affected}) \\
& P(I_1 = i_1,\ I_2 = i_2 \mid \text{both affected})\}
\end{aligned} \tag{4.10}
$$

$$
\begin{aligned}
= \; & \sum_{i_1=0}^{2} \sum_{i_2=0}^{2} P(X_{1j} = x, X_{2j} = y \mid I_1 = i_1) P(X_{3j} = u, X_{4j} = v \mid I_2 = i_2) \\
& P(I_1 = i_1,\ I_2 = i_2 \mid \text{both affected})
\end{aligned} \tag{4.11}
$$

Table 4.5. Conditional distribution of X_{1j} and X_{2j} given I_1.

i	x	y	$P(X_{1j}=x,\ X_{2j}=y \mid I_1=i)$
2	0	0	$(1-\Psi_1^2)(1-\Psi_2^2)$
	1	0	$\Psi_1^2(1-\Psi_2^2)$
	0	1	$(1-\Psi_1^2)\Psi_2^2$
	1	1	$\Psi_1^2\Psi_2^2$
1	0	0	$(1-\Psi_1+\Psi_1^2)(1-\Psi_2+\Psi_2^2)$
	1	0	$\Psi_1(1-\Psi_1)(1-\Psi_2+\Psi_2^2)$
	0	1	$(1-\Psi_1+\Psi_1^2)\Psi_2(1-\Psi_2)$
	1	1	$\Psi_1(1-\Psi_1)\Psi_2(1-\Psi_2)$
0	0	0	$(2\Psi_1-\Psi_1^2)(2\Psi_2-\Psi_2^2)$
	1	0	$(1-\Psi_1)^2(2\Psi_2-\Psi_2^2)$
	0	1	$(2\Psi_1-\Psi_1^2)(1-\Psi_2)^2$
	1	1	$(1-\Psi_1)^2(1-\Psi_2)^2$

The probability $P(X_{1j}=x,\ X_{2j}=y \mid I_1=i)$ in general case, i.e. gene is anywhere in between 2 markers, can be derived from Table 2.3, and is given in Table 4.5. Where $\Psi_j=\theta_j^2+(1-\theta_j)^2$, j=1,2, θ_1 and θ_2 are the recombination fraction between gene and marker 1, and between gene and marker 2, respectively. The probability $P(X_{3j}=x,\ X_{4j}=y \mid I_2=i)$ is same except using different θs.

The probability $P(I_1=i_1,\ I_2=i_2 \mid \text{both affected})$ is equal to,

$$\frac{P(I_1=i_1,\ I_2=i_2,\ \text{both affected})}{P(\text{both affected})}$$

$$= \frac{\sum_{E_1}\sum_{E_2}\sum_{PG_1}\sum_{PG_2} P(I_1=i_1,\ I_2=i_2,\ E_1,\ E_2,\ PG_1,\ PG_2,\ \text{both affected})}{P(\text{both affected})}$$

$$= \{\sum_{E_1}\sum_{E_2}\sum_{PG_1}\sum_{PG_2} P(\text{both affected}|I_1=i,\ I_2=i_2\ E_1,\ E_2,\ PG_1,\ PG_2)$$

$$P(I_1 = i_1,\ I_2 = i_2\ E_1,\ E_2,\ PG_1,\ PG_2)\}/P(\text{both affected})$$

$$= \frac{\sum_{E_1}\sum_{E_2}\sum_{PG_1}\sum_{PG_2} P(\text{both affected}|E_1,\ E_2)P(I_1 = i_1,\ I_2 = i_2\ E_1,\ E_2,\ PG_1,\ PG_2)}{P(\text{both affected})}$$

$$= \frac{\{\sum_{E_1}\sum_{E_2}\sum_{PG_1}\sum_{PG_2} P(\text{both affected}|E_1,\ E_2)P(I_1 = i_1,\ E_1,\ PG_1)P(I_2 = i_2,\ E_2,\ PG_2)\}}{P(\text{both affected})}$$

$$= \frac{1}{P(\text{both affected})}\{\sum_{E_1}\sum_{E_2}\sum_{PG_1}\sum_{PG_2} P(\text{both affected}|E_1,\ E_2)P(I_1 = i_1 \mid E_1,\ PG_1)$$

$$P(E_1 \mid PG_1)P(PG_1)P(I_2 = i_2 \mid E_2,\ PG_2)P(E_2 \mid PG_2)P(PG_2)\}$$

Because both sibs are affected, hence when $k_0 = 0$ the following restriction on summing over E_1 and E_2 are applied; if $E_1 \neq (1,1)$ and $E_2 \neq (1,1)$ then $(E_1,\ E_2)$ must be ((1,0), (0,1)) or ((0,1),(1,0)). $P(\text{both affected}|E_1,\ E_2)$ is given in Table 4.6. The probabilities $P(\text{both affected}) = \sum_{i=0}^{2}\sum_{j=0}^{2} P(I_1 = i, I_2 = j, \text{both affected})$. $P(I_1 = i,\ I_2 = j,$ both affected) are given as follows. The rest probabilities were given in Table 4.1. Therefore, $P(I_1 = i_1,\ I_2 = i_2 \mid$ both affected) can be found.

Let $p_1 = P(PG = 1)$, $p_2 = P(PG = 2)$, $p_3 = P(PG = 3)$, and $p_{456} = P(PG =$ 4, 5, or 6), Then,

$$P(I_1 = 0, I_2 = 0, BA)$$

$$= 4k_0\lambda^2\left(\frac{1}{4}p_2\frac{1}{2} + \frac{3}{16}p_3\frac{1}{3}\right)\left(\frac{9}{16}p_3\frac{2}{9} + \frac{1}{4}p_{456}\right)$$

$$+2\lambda^2\left(\frac{1}{4}p_2\frac{1}{2} + \frac{3}{16}p_3\frac{1}{3}\right)\left(\frac{1}{4}p_2\frac{1}{2} + \frac{3}{16}p_3\frac{1}{3}\right)$$

$$+4k_0\lambda^2\left(\frac{1}{4}p_2\frac{1}{2} + \frac{3}{16}p_3\frac{1}{3}\right)^2$$

$$+k_0^2\lambda^2\left(\frac{9}{16}p_3\frac{2}{9} + \frac{1}{4}p_{456}\right)^2$$

$$= \frac{1}{128}\lambda^2(2p_2 + p_3)^2$$

Table 4.6. $P(\text{both affected}|E_1,\ E_2)$ and possible trait IBD given E_1 and E_2.

E_1	E_2	$P(\text{BA}\|E_1,\ E_2)$	possible trait IBD score, (I_1, I_2), given E_1 and E_2
(1,1)	(0,0)	λ^2	(2,0), (2,1), (2,2)
(1,1)	(0,1)	$2\lambda^2$	(2,0), (2,1)
(1,1)	(1,0)	$2\lambda^2$	(2,0), (2,1)
(1,1)	(1,1)	$4\lambda^2$	(2,2)
(1,0)	(0,0)	$k_0\lambda^2$	(1,0),(1,1),(1,2),(0,0),(0,1),(0,2)
(1,0)	(0,1)	λ^2	(1,0),(1,1),(0,0),(0,1)
(1,0)	(1,0)	$2k_0\lambda^2$	(1,0),(1,1),(0,0),(0,1)
(1,0)	(1,1)	$2\lambda^2$	(1,2),(0,2)
(0,1)	(0,0)	$k_0\lambda^2$	(1,0),(1,1),(1,2),(0,0),(0,1),(0,2)
(0,1)	(0,1)	$2k_0\lambda^2$	(1,0),(1,1),(0,0),(0,1)
(0,1)	(1,0)	λ^2	(1,0),(1,1),(0,0),(0,1)
(0,1)	(1,1)	$2\lambda^2$	(1,2),(0,2)
(0,0)	(0,0)	$k_0^2\lambda^2$	(2,0),(2,1),(2,2),(1,0),(1,1),(1,2),(0,0),(0,1),(0,2)
(0,0)	(0,1)	$k_0\lambda^2$	(2,0),(2,1),(1,0),(1,1),(0,0),(0,1)
(0,0)	(1,0)	$k_0\lambda^2$	(2,0),(2,1),(1,0),(1,1),(0,0),(0,1)
(0,0)	(1,1)	λ^2	(2,2),(1,2),(0,2)

$$+\frac{1}{64}k_0\lambda^2(2p_2+p_3)(2p_2+3p_3+4p_{456})$$

$$+\frac{1}{64}k_0^2\lambda^2(p_3+2p_{456})^2$$

$$P(I_1=1,I_2=0,BA)=P(I_1=2,I_2=1,BA)$$

$$\begin{aligned}
&= 2k_0\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{2}{3}\right)\left(\frac{9}{16}p_3\frac{2}{9}+\frac{1}{4}p_{456}\right)\\
&\quad+2\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{2}{3}\right)\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{1}{3}\right)\\
&\quad+4k_0\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{2}{3}\right)\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{1}{3}\right)\\
&\quad+2k_0\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{4}{9}+\frac{1}{2}p_{456}p_{456}\right)\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{1}{3}\right)\\
&\quad+k_0^2\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{4}{9}+\frac{1}{2}p_{456}p_{456}\right)\left(\frac{9}{16}p_3\frac{2}{9}+\frac{1}{4}p_{456}\right)^2\\
&= \frac{1}{64}\lambda^2(p_2+p_3)(2p_2+p_3)\\
&\quad+\frac{1}{16}k_0\lambda^2(p_2+p_3)(p_2+p_3+p_{456})\\
&\quad+\frac{1}{64}k_0\lambda^2(p_2+2p_3+4p_{456})(2p_2+p_3)\\
&\quad+\frac{1}{64}k_0^2\lambda^2(p_2+2p_3+4p_{456})(p_3+2p_{456})
\end{aligned}$$

$$P(I_1=2,I_2=0,BA)=P(I_1=0,I_2=2,BA)$$

$$\begin{aligned}
&= \lambda^2\left(p_1+\frac{1}{4}p_2+\frac{1}{16}p_3\right)\left(\frac{9}{16}p_3\frac{2}{9}+\frac{1}{4}p_{456}\right)\\
&\quad+4\lambda^2\left(p_1+\frac{1}{4}p_2+\frac{1}{16}p_3\right)\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{1}{3}\right)
\end{aligned}$$

$$
\begin{aligned}
&+k_0^2\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{2}{9}+\frac{1}{4}p_{456}\right)\left(\frac{9}{16}p_3\frac{2}{9}+\frac{1}{4}p_{456}\right)\\
&+2k_0\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{3}{9}+\frac{1}{4}p_{456}\right)\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{1}{3}\right)\\
=\ &\frac{1}{128}\lambda^2(16p_1+4p_2+p_3)(4p_2+3p_3+2p_{456})\\
&+\frac{1}{128}k_0\lambda^2(4p_2+3p_3+4p_{456})(2p_2+p_3)\\
&+\frac{1}{128}k_0^2\lambda^2(2p_2+3p_3+4p_{456})(p_3+2p_{456})
\end{aligned}
$$

$$
\begin{aligned}
&P(I_1=1,I_2=1,BA)\\
=\ &4k_0\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{2}{3}\right)\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{4}{9}+\frac{1}{2}p_{456}\right)\\
&+2\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{2}{3}\right)^2\\
&+4k_0\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{2}{3}\right)^2\\
&+k_0^2\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{4}{9}+\frac{1}{2}p_{456}\right)^2\\
=\ &\frac{1}{32}\lambda^2(p_2+p_3)^2\\
&+\frac{1}{16}k_0\lambda^2(p_2+p_3)(2p_2+3p_3+4p_{456})\\
&+\frac{1}{64}k_0^2\lambda^2(p_2+2p_3+4p_{456})^2
\end{aligned}
$$

$$
\begin{aligned}
&P(I_1=2,I_2=1,BA)=P(I_1=1,I_2=2,BA)\\
=\ &\lambda^2\left(p_1+\frac{1}{4}p_2+\frac{1}{16}p_3\right)\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{4}{9}+\frac{1}{2}p_{456}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+4\lambda^2\left(p_1+\frac{1}{4}p_2+\frac{1}{16}p_3\right)\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{2}{3}\right)\\
&+k_0^2\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{3}{9}+\frac{1}{4}p_{456}\right)\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{4}{9}+\frac{1}{2}p_{456}\right)\\
&+2k_0\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{3}{9}+\frac{1}{4}p_{456}\right)\left(\frac{1}{4}p_2\frac{1}{2}+\frac{3}{16}p_3\frac{2}{3}\right)\\
=\ &\frac{1}{128}\lambda^2(16p_1+4p_2+p_3)(5p_2+7p_3+4p_{456})\\
&+\frac{1}{64}k_0\lambda^2(2p_2+3p_3+4p_{456})(p_2+p_3)\\
&+\frac{1}{128}k_0^2\lambda^2(2p_2+3p_3+4p_{456})(p_2+2p_3+4p_{456})
\end{aligned}
$$

$$
\begin{aligned}
&P\,(I_1=2,I_2=2,BA)\\
=\ &2\lambda^2\left(p_1+\frac{1}{4}p_2+\frac{1}{16}p_3\right)\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{4}{9}+\frac{1}{4}p_{456}\right)\\
&+4\lambda^2\left(p_1+\frac{1}{4}p_2+\frac{1}{16}p_3\right)^2\\
&+k_0^2\lambda^2\left(\frac{1}{4}p_2\frac{1}{2}+\frac{9}{16}p_3\frac{3}{9}+\frac{1}{4}p_{456}\right)^2\\
=\ &\frac{1}{128}\lambda^2\,(16p_1+4p_2+p_3)\,(32p_1+10p_2+5p_3+4p_{456})\\
&+\frac{1}{256}k_0^2\lambda^2\,(2p_2+3p_3+4p_{456})^2
\end{aligned}
$$

Therefore, the joint distribution of X_{1j} and X_{2j} and X_{3j} and X_{4j} conditional on both sibs are affected, p_{xyuv}, can be found. Even though p_{xyuv} is known, it takes too much time to compute the exact joint distribution of ${}_1S_1$, ${}_1S_2$, ${}_1S_3$ and ${}_1S_4$ conditional on both sibs are affected using the same method we used in the chapter 3 for joint

distribution of ${}_lS_0$ and ${}_lS_1$. Therefore, Monte Carlo integration were used to find the conditional joint distribution of ${}_lS_1$, ${}_lS_2$, ${}_lS_3$ and ${}_1S_4$.

4.3.3 Simulation under More Realistic Assumptions

For the same reason as in the simple disease case, we also performed simulations for an independent model and a Markov chain model without combining Stage I and Stage II data. The total length of chromosomes is 3300 cM and divided into 22 chromosomes. The number of genes is two. The combinations of parameters are: resource $N = 2000$, 5000, 7500, 10,000, and 20,000; number of ASP $n = 50$, 100, and 150; population frequency (P.F.) of disease gene = 0.01, 0.2, and 0.5; relative risk, k_0, of being affected of carrying no gene vs. carrying two recessive genes at one locus is equal to 0, 0.05, 0.1 and 1.0. Map densities used are 5 cM, 10 cM, and 20 cM. Threshold k from 1 to n_1, and n_1 from 5 to $n-5$ with increment 5. The simulations were conducted similar to the simulation in chapter 3. We simulated 22 pairs of chromosomes instead of one; two genes instead of one. The details of simulation are as follows.

1. Reading in parameters, resource N, P.F., k_0, number of ASP n, and map, n_1, and k of the design.

2. Generate gene location from uniform distribution(0,3300/22) twice, first one for the gene on chromosome 1, and second one for the gene on chromosome 2. Then follow the rest of step 2 on page 61. For those chromosomes were not chosen, set l=0. Generate joint trait IBD for two loci.

3. Follow step 3 and 4 on page 61 for every chromosome.

4. Repeat step 2 and 3 n_1 times and then calculated ${}_1S_i$.

5. Check if any of ${}_1S_1$, ${}_1S_2$, ${}_1S_3$, and ${}_1S_4 > k$. If yes, then calculate N_2 according to the resource constraint, choose markers whose ${}_2S_i > k$ for Stage II study,

and go to the next step. If not, record the detection as a failure and go back to step 2.

6. The X_{ij}s in Stage II are generated the same way as those in Stage I, except only the chosen markers are used and repeat N_2 times instead of n_1.

7. Check whether ${}_2S_1$ or ${}_2S_2$, or ${}_2S_3$ or ${}_2S_4$, according to which one passed Stage I, are the largest among all ${}_2S_i$s in Stage II. If any one of them is the largest, it is a success; otherwise it is a failure. Go back to step 2 until enough simulation has been done.

The programs were compiled and ran in the same computing environment mentioned in chapter 3.

4.3.4 Results

Based on analytic computation and simulation, the design with the highest probability of finding the correct marker(s) was identified for a disease with two unlinked recessive genes with additive penetrance. These optimal designs are given in Table 4.7.

In the tables, the P.F. column shows the population frequency of the disease gene; the k_0 columns, the relative risk; the map columns, the map density; the k column, the threshold for passing Stage I for Stage II study; the n_1 columns, the number of ASP used in Stage I; the F2 column, the probability of locating the right marker by the best two-stage design obtained by analytic formula; the I.1, I.2, and I.3 columns show the probabilities, obtained by simulation with independent assumption, of finding the markers next to the genes, within two intervals, and within three intervals, respectively, by the two-stage design. M.1, M.2, and M.3 columns show the probabilities corresponding to I.1, I.2, and I.3 obtained by simulation with Markov chain assumption without combining first- and second-stage data; The F1 column

Table 4.7. Optimal resource allocation in two-stage genome search for rare recessive gene, assumed two genes.

Resource N=2,000, ASP=50															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	20 cM	2	5	0.857	0.852	0.853	0.854	0.838	0.855	0.859	20 cM	12	0.384	0.473*
0.01	0.05	20 cM	5	10	0.023	0.019	0.020	0.021	0.020	0.022	0.024	20 cM	12	0013	0.010
0.01	0.1	20 cM	5	10	0.020	0.018	0.019	0.019	0.019	0.021	0.022	20 cM	12	0.012	0.008
0.01	1.0	20 cM	5	10	0.019	0.021	0.022	0.023	0.018	0.021	0.022	20 cM	12	0.011	0.008
0.20	0.0	20 cM	2	5	0.804	0.798	0.799	0.800	0.777	0.799	0.804	20 cM	12	0.332	0.472*
0.20	0.05	20 cM	2	5	0.590	0.585	0.588	0.591	0.561	0.585	0.592	20 cM	12	0.199	0.391
0.20	0.1	20 cM	2	5	0.361	0.345	0.350	0.355	0.334	0.359	0.367	20 cM	12	0.111	0.250
0.20	1.0	20 cM	5	10	0.021	0.018	0.019	0.019	0.019	0.022	0.023	20 cM	12	0.013	0.009
0.50	0.0	20 cM	2	5	0.844	0.837	0.838	0.838	0.820	0.836	0.840	20 cM	12	0.380	0.464*
0.50	0.05	20 cM	2	5	0.804	0.804	0.805	0.806	0.780	0.798	0.802	20 cM	12	0.340	0.464*
0.50	0.1	20 cM	2	5	0.761	0.748	0.751	0.752	0.733	0.752	0.756	20 cM	12	0.303	0.458*
0.50	1.0	20 cM	2	5	0.142	0.133	0.138	0.143	0.129	0.148	0.156	20 cM	12	0.049	0.093

Table 4.7–continued

Resource N=2,000, ASP=100															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	20 cM	2	5	0.891	0.876	0.876	0.877	0.857	0.869	0.872	20 cM	12	0.384	0.507*
0.01	0.05	20 cM	5	10	0.024	0.022	0.022	0.023	0.021	0.024	0.026	20 cM	12	0.013	0.011
0.01	0.1	20 cM	5	10	0.021	0.018	0.019	0.019	0.019	0.021	0.021	20 cM	12	0.012	0.009
0.01	1.0	20 cM	5	10	0.020	0.020	0.020	0.021	0.017	0.020	0.021	20 cM	12	0.011	0.009
0.20	0.0	20 cM	2	5	0.849	0.836	0.837	0.837	0.812	0.830	0.834	20 cM	12	0.332	0.517*
0.20	0.05	20 cM	2	5	0.669	0.629	0.632	0.633	0.605	0.628	0.633	20 cM	12	0.199	0.470*
0.20	0.1	20 cM	2	5	0.442	0.402	0.407	0.409	0.394	0.418	0.426	20 cM	12	0.111	0.330
0.20	1.0	20 cM	5	10	0.022	0.019	0.020	0.022	0.023	0.026	0.027	20 cM	12	0.013	0.010
0.50	0.0	20 cM	2	5	0.875	0.861	0.862	0.862	0.851	0.865	0.867	20 cM	12	0.380	0.495*
0.50	0.05	20 cM	2	5	0.844	0.837	0.837	0.838	0.819	0.833	0.836	20 cM	12	0.340	0.504*
0.50	0.1	20 cM	2	5	0.810	0.789	0.790	0.791	0.772	0.788	0.791	20 cM	12	0.303	0.507*
0.50	1.0	20 cM	2	5	0.181	0.156	0.162	0.167	0.160	0.179	0.186	20 cM	12	0.049	0.132

Table 4.7–continued

Resource N=2,000, ASP=150															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	20 cM	2	5	0.891	0.881	0.881	0.882	0.864	0.878	0.880	20 cM	12	0.384	0.507*
0.01	0.05	20 cM	5	10	0.025	0.022	0.024	0.024	0.021	0.023	0.024	20 cM	12	0.013	0.011
0.01	0.1	20 cM	5	10	0.021	0.021	0.022	0.023	0.020	0.023	0.024	20 cM	12	0.012	0.009
0.01	1.0	20 cM	3	5	0.020	0.017	0.018	0.018	0.019	0.021	0.021	20 cM	12	0.011	0.009
0.20	0.0	20 cM	2	5	0.849	0.832	0.834	0.834	0.810	0.825	0.829	20 cM	12	0.332	0.517*
0.20	0.05	20 cM	2	5	0.670	0.637	0.639	0.641	0.614	0.636	0.640	20 cM	12	0.199	0.471*
0.20	0.1	20 cM	2	5	0.443	0.414	0.419	0.423	0.388	0.410	0.417	20 cM	12	0.111	0.331
0.20	1.0	20 cM	5	10	0.022	0.020	0.020	0.021	0.019	0.021	0.022	20 cM	12	0.013	0.010
0.50	0.0	20 cM	2	5	0.875	0.863	0.863	0.864	0.847	0.861	0.863	20 cM	12	0.380	0.495*
0.50	0.05	20 cM	2	5	0.844	0.828	0.829	0.830	0.806	0.820	0.821	20 cM	12	0.340	0.504*
0.50	0.1	20 cM	2	5	0.810	0.790	0.792	0.792	0.762	0.778	0.780	20 cM	12	0.303	0.507*
0.50	1.0	20 cM	2	5	0.182	0.160	0.164	0.169	0.156	0.175	0.182	20 cM	12	0.049	0.133

Table 4.7–continued

Resource N=5,000, ASP=50															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	10 cM	4	10	0.985	0.987	0.987	0.987	0.946	0.983	0.986	20 cM	30	0.851	0.135
0.01	0.05	20 cM	5	10	0.023	0.021	0.022	0.023	0.020	0.023	0.023	20 cM	30	0.017	0.006
0.01	0.1	20 cM	5	10	0.020	0.019	0.019	0.020	0.020	0.022	0.023	20 cM	30	0.015	0.005
0.01	1.0	20 cM	5	10	0.019	0.018	0.019	0.020	0.018	0.021	0.022	20 cM	30	0.015	0.005
0.20	0.0	10 cM	4	10	0.964	0.964	0.965	0.965	0.914	0.956	0.965	20 cM	30	0.789	0.176
0.20	0.05	10 cM	4	10	0.795	0.784	0.786	0.787	0.721	0.786	0.799	20 cM	30	0.541	0.254
0.20	0.1	10 cM	4	10	0.489	0.487	0.489	0.491	0.432	0.492	0.510	20 cM	30	0.305	0.184
0.20	1.0	20 cM	7	15	0.021	0.020	0.021	0.022	0.019	0.021	0.022	20 cM	30	0.017	0.005
0.50	0.0	10 cM	4	10	0.979	0.977	0.978	0.978	0.942	0.975	0.979	20 cM	30	0.843	0.136
0.50	0.05	10 cM	4	10	0.962	0.962	0.962	0.962	0.916	0.954	0.959	20 cM	30	0.794	0.167
0.50	0.1	10 cM	4	10	0.938	0.929	0.929	0.930	0.885	0.931	0.939	20 cM	30	0.741	0.197
0.50	1.0	20 cM	1	5	0.169	0.156	0.165	0.174	0.154	0.183	0.193	20 cM	30	0.113	0.056

Table 4.7–continued.

Resource N=5,000, ASP=100															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	20 cM	7	20	0.998	0.997	0.997	0.997	0.994	0.997	0.997	20 cM	30	0.851	0.147
0.01	0.05	20 cM	10	25	0.026	0.022	0.023	0.025	0.021	0.024	0.026	20 cM	30	0.017	0.009
0.01	0.1	20 cM	10	25	0.021	0.019	0.020	0.021	0.019	0.023	0.024	20 cM	30	0.015	0.006
0.01	1.0	20 cM	5	10	0.020	0.018	0.019	0.020	0.020	0.021	0.022	20 cM	30	0.015	0.006
0.20	0.0	20 cM	7	20	0.994	0.992	0.992	0.993	0.985	0.991	0.991	20 cM	30	0.789	0.205
0.20	0.05	20 cM	5	15	0.930	0.920	0.921	0.922	0.899	0.918	0.922	20 cM	30	0.541	0.388
0.20	0.1	20 cM	5	15	0.713	0.693	0.696	0.700	0.663	0.691	0.699	20 cM	30	0.305	0.408*
0.20	1.0	20 cM	10	25	0.024	0.020	0.023	0.024	0.026	0.030	0.030	20 cM	30	0.017	0.007
0.50	0.0	20 cM	7	20	0.997	0.996	0.996	0.996	0.991	0.994	0.995	20 cM	30	0.843	0.154
0.50	0.05	20 cM	7	20	0.993	0.991	0.991	0.991	0.987	0.992	0.992	20 cM	30	0.794	0.199
0.50	0.1	20 cM	7	20	0.987	0.985	0.986	0.986	0.975	0.983	0.984	20 cM	30	0.741	0.246
0.50	1.0	20 cM	5	15	0.297	0.271	0.278	0.284	0.274	0.302	0.313	20 cM	30	0.113	0.184

Table 4.7–continued.

Resource N=5,000, ASP=150															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	20 cM	7	20	0.998	0.998	0.998	0.998	0.993	0.997	0.997	20 cM	30	0.851	0.147
0.01	0.05	20 cM	10	25	0.027	0.022	0.024	0.025	0.024	0.027	0.029	20 cM	30	0.017	0.010
0.01	0.1	20 cM	8	20	0.022	0.021	0.023	0.025	0.021	0.024	0.026	20 cM	30	0.015	0.007
0.01	1.0	20 cM	10	25	0.021	0.020	0.021	0.022	0.018	0.020	0.022	20 cM	30	0.015	0.006
0.20	0.0	20 cM	7	20	0.995	0.991	0.992	0.992	0.985	0.991	0.992	20 cM	30	0.789	0.206
0.20	0.05	20 cM	5	15	0.943	0.927	0.928	0.928	0.905	0.922	0.925	20 cM	30	0.541	0.402*
0.20	0.1	20 cM	5	15	0.753	0.708	0.711	0.713	0.682	0.712	0.718	20 cM	30	0.305	0.448*
0.20	1.0	20 cM	10	25	0.025	0.021	0.022	0.023	0.021	0.026	0.028	20 cM	30	0.017	0.008
0.50	0.0	20 cM	7	20	0.997	0.997	0.997	0.997	0.992	0.995	0.996	20 cM	30	0.843	0.154
0.50	0.05	20 cM	7	20	0.994	0.993	0.993	0.993	0.988	0.992	0.993	20 cM	30	0.794	0.200
0.50	0.1	20 cM	7	20	0.989	0.986	0.986	0.986	0.977	0.984	0.985	20 cM	30	0.741	0.248
0.50	1.0	20 cM	5	15	0.333	0.294	0.299	0.303	0.283	0.309	0.319	20 cM	30	0.113	0.220

Table 4.7–continued.

Resource N=7,500, ASP=50															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	10 cM	3	10	0.992	0.992	0.992	0.992	0.956	0.988	0.992	20 cM	45	0.966	0.027
0.01	0.05	20 cM	5	10	0.023	0.020	0.022	0.023	0.021	0.024	0.025	20 cM	45	0.019	0.004
0.01	0.1	20 cM	5	10	0.020	0.016	0.017	0.017	0.019	0.020	0.021	20 cM	45	0.017	0.004
0.01	1.0	20 cM	5	10	0.019	0.018	0.019	0.020	0.019	0.021	0.022	20 cM	45	0.016	0.003
0.20	0.0	10 cM	3	10	0.981	0.980	0.980	0.980	0.934	0.977	0.982	20 cM	45	0.937	0.044
0.20	0.05	10 cM	1	5	0.851	0.837	0.838	0.840	0.755	0.825	0.842	20 cM	45	0.748	0.104
0.20	0.1	10 cM	1	5	0.548	0.527	0.530	0.533	0.474	0.545	0.569	20 cM	45	0.462	0.086
0.20	1.0	20 cM	7	15	0.021	0.019	0.020	0.021	0.019	0.020	0.022	20 cM	45	0.018	0.003
0.50	0.0	10 cM	3	10	0.991	0.990	0.990	0.990	0.953	0.987	0.991	20 cM	45	0.961	0.030
0.50	0.05	10 cM	3	10	0.982	0.983	0.983	0.983	0.939	0.977	0.983	20 cM	45	0.939	0.043
0.50	0.1	10 cM	3	10	0.968	0.965	0.965	0.966	0.916	0.960	0.968	20 cM	45	0.908	0.060
0.50	1.0	10 cM	1	5	0.175	0.167	0.171	0.176	0.155	0.196	0.210	20 cM	45	0.167	0.008

Table 4.7–continued.

Resource N=7,500, ASP=100															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	20 cM	4	15	0.999	0.999	0.999	0.999	0.997	0.999	0.999	20 cM	45	0.966	0.034
0.01	0.05	20 cM	10	25	0.026	0.021	0.023	0.024	0.021	0.026	0.027	20 cM	45	0.019	0.007
0.01	0.1	20 cM	10	25	0.021	0.017	0.019	0.020	0.019	0.023	0.024	20 cM	45	0.017	0.005
0.01	1.0	20 cM	12	30	0.020	0.017	0.018	0.019	0.018	0.020	0.020	20 cM	45	0.016	0.004
0.20	0.0	20 cM	4	15	0.998	0.997	0.997	0.998	0.990	0.995	0.996	20 cM	45	0.937	0.061
0.20	0.05	20 cM	4	15	0.961	0.952	0.953	0.954	0.928	0.947	0.949	20 cM	45	0.748	0.213
0.20	0.1	20 cM	4	15	0.772	0.747	0.751	0.754	0.713	0.747	0.755	20 cM	45	0.462	0.310
0.20	1.0	20 cM	12	30	0.024	0.020	0.021	0.022	0.022	0.025	0.026	20 cM	45	0.018	0.006
0.50	0.0	20 cM	4	15	0.999	0.999	0.999	0.999	0.995	0.997	0.998	20 cM	45	0.961	0.038
0.50	0.05	20 cM	6	20	0.998	0.998	0.998	0.998	0.995	0.998	0.999	20 cM	45	0.939	0.059
0.50	0.1	20 cM	4	15	0.995	0.994	0.995	0.995	0.985	0.991	0.991	20 cM	45	0.908	0.088
0.50	1.0	20 cM	4	15	0.326	0.300	0.309	0.318	0.292	0.327	0.341	20 cM	45	0.167	0.159

Table 4.7–continued.

Resource N=7,500, ASP=150															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	20 cM	4	15	1.000	0.999	0.999	0.999	0.997	0.999	0.999	20 cM	45	0.966	0.034
0.01	0.05	20 cM	13	35	0.027	0.021	0.023	0.024	0.027	0.031	0.033	20 cM	45	0.019	0.008
0.01	0.1	20 cM	10	25	0.022	0.017	0.018	0.019	0.022	0.025	0.027	20 cM	45	0.017	0.006
0.01	1.0	20 cM	15	40	0.021	0.019	0.021	0.021	0.023	0.025	0.026	20 cM	45	0.016	0.005
0.20	0.0	20 cM	6	20	0.999	0.999	0.999	0.999	0.997	0.998	0.999	20 cM	45	0.937	0.063
0.20	0.05	20 cM	8	25	0.987	0.979	0.979	0.979	0.973	0.980	0.981	20 cM	45	0.748	0.239
0.20	0.1	20 cM	8	25	0.874	0.849	0.850	0.852	0.826	0.850	0.855	20 cM	45	0.462	0.412
0.20	1.0	20 cM	13	35	0.025	0.020	0.021	0.023	0.023	0.026	0.028	20 cM	45	0.018	0.007
0.50	0.0	20 cM	4	15	0.999	0.999	0.999	0.999	0.995	0.998	0.998	20 cM	45	0.961	0.038
0.50	0.05	20 cM	6	20	0.999	0.999	0.999	0.999	0.998	0.999	0.999	20 cM	45	0.939	0.060
0.50	0.1	20 cM	10	30	0.999	0.999	0.999	0.999	0.997	0.998	0.999	20 cM	45	0.908	0.091
0.50	1.0	20 cM	8	25	0.427	0.392	0.399	0.406	0.381	0.411	0.423	20 cM	45	0.167	0.260

Table 4.7–continued.

Resource N=10,000, ASP=50															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	5 cM	4	10	0.994	0.995	0.995	0.995	0.887	0.974	0.989	20 cM	50	0.980	0.015
0.01	0.05	20 cM	5	10	0.023	0.016	0.017	0.018	0.020	0.023	0.023	20 cM	50	0.019	0.004
0.01	0.1	20 cM	5	10	0.020	0.017	0.018	0.018	0.018	0.020	0.021	20 cM	50	0.017	0.003
0.01	1.0	20 cM	5	10	0.019	0.020	0.020	0.021	0.018	0.020	0.021	20 cM	50	0.017	0.003
0.20	0.0	10 cM	2	10	0.983	0.978	0.979	0.979	0.923	0.968	0.976	20 cM	50	0.959	0.024
0.20	0.05	10 cM	1	5	0.855	0.851	0.853	0.853	0.790	0.854	0.870	20 cM	50	0.797	0.058
0.20	0.1	10 cM	1	5	0.552	0.540	0.544	0.548	0.498	0.563	0.587	10 cM	30	0.353	0.199
0.20	1.0	20 cM	7	15	0.021	0.018	0.019	0.019	0.023	0.025	0.026	20 cM	50	0.019	0.003
0.50	0.0	10 cM	2	10	0.992	0.989	0.989	0.990	0.942	0.981	0.986	20 cM	50	0.977	0.015
0.50	0.05	10 cM	2	10	0.983	0.977	0.977	0.978	0.921	0.969	0.975	20 cM	50	0.961	0.022
0.50	0.1	10 cM	2	10	0.968	0.962	0.962	0.962	0.901	0.953	0.960	20 cM	50	0.936	0.031
0.50	1.0	10 cM	1	5	0.176	0.165	0.171	0.175	0.161	0.200	0.219	20 cM	50	0.185	-0.009

Table 4.7–continued.

Resource N=10,000, ASP=100															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	20 cM	2	10	1.000	0.999	0.999	0.999	0.998	1.000	1.000	20 cM	60	0.993	0.006
0.01	0.05	20 cM	10	25	0.026	0.021	0.023	0.024	0.021	0.024	0.026	20 cM	60	0.020	0.006
0.01	0.1	20 cM	10	25	0.021	0.020	0.021	0.022	0.020	0.023	0.025	20 cM	60	0.017	0.004
0.01	1.0	20 cM	12	30	0.020	0.019	0.020	0.021	0.019	0.021	0.022	20 cM	60	0.017	0.003
0.20	0.0	10 cM	3	10	0.999	0.999	0.999	0.999	0.987	0.998	0.998	20 cM	60	0.983	0.016
0.20	0.05	10 cM	7	20	0.985	0.985	0.986	0.986	0.945	0.978	0.982	20 cM	60	0.870	0.115
0.20	0.1	10 cM	7	20	0.842	0.833	0.834	0.835	0.765	0.822	0.835	20 cM	60	0.597	0.245
0.20	1.0	20 cM	12	30	0.024	0.018	0.019	0.021	0.022	0.024	0.025	20 cM	60	0.019	0.004
0.50	0.0	20 cM	2	10	0.999	0.999	0.999	0.999	0.997	0.999	0.999	20 cM	60	0.991	0.008
0.50	0.05	10 cM	5	15	0.999	0.999	0.999	0.999	0.992	0.999	1.000	20 cM	60	0.984	0.016
0.50	0.1	10 cM	7	20	0.999	0.999	0.999	0.999	0.988	0.999	0.999	20 cM	60	0.971	0.028
0.50	1.0	10 cM	5	15	0.354	0.347	0.351	0.354	0.308	0.355	0.373	20 cM	60	0.222	0.133

Table 4.7–continued.

Resource N=10,000, ASP=150															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	20 cM	2	10	1.000	0.999	0.999	0.999	0.998	1.000	1.000	20 cM	60	0.993	0.007
0.01	0.05	20 cM	13	35	0.027	0.025	0.027	0.029	0.023	0.026	0.028	20 cM	60	0.020	0.007
0.01	0.1	20 cM	10	25	0.022	0.020	0.021	0.022	0.019	0.022	0.024	20 cM	60	0.017	0.005
0.01	1.0	20 cM	15	40	0.021	0.018	0.019	0.020	0.021	0.024	0.025	20 cM	60	0.017	0.004
0.20	0.0	20 cM	2	10	0.999	0.999	0.999	0.999	0.995	0.998	0.999	20 cM	60	0.983	0.016
0.20	0.05	20 cM	9	30	0.994	0.993	0.993	0.993	0.988	0.994	0.994	20 cM	60	0.870	0.124
0.20	0.1	20 cM	9	30	0.910	0.894	0.895	0.896	0.879	0.900	0.905	20 cM	60	0.597	0.314
0.20	1.0	20 cM	13	35	0.025	0.022	0.024	0.026	0.023	0.027	0.029	20 cM	60	0.019	0.006
0.50	0.0	20 cM	2	10	1.000	1.000	1.000	1.000	0.998	1.000	1.000	20 cM	60	0.991	0.008
0.50	0.05	20 cM	2	10	0.999	0.998	0.998	0.998	0.996	0.998	0.998	20 cM	60	0.984	0.015
0.50	0.1	20 cM	4	15	0.999	0.999	0.999	0.999	0.996	0.998	0.998	20 cM	60	0.971	0.029
0.50	1.0	20 cM	9	30	0.464	0.444	0.449	0.456	0.426	0.459	0.469	20 cM	60	0.222	0.242

Table 4.7–continued.

Resource N=20,000, ASP=50															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	5 cM	2	10	0.998	0.998	0.998	0.998	0.885	0.975	0.992	10 cM	50	0.999	-0.000
0.01	0.05	20 cM	5	10	0.023	0.022	0.023	0.024	0.020	0.023	0.025	20 cM	50	0.019	0.004
0.01	0.1	20 cM	5	10	0.020	0.019	0.019	0.020	0.017	0.019	0.020	20 cM	50	0.017	0.003
0.01	1.0	20 cM	5	10	0.019	0.017	0.018	0.018	0.016	0.018	0.019	20 cM	50	0.017	0.003
0.20	0.0	5 cM	2	10	0.992	0.991	0.991	0.991	0.853	0.956	0.983	10 cM	50	0.995	-0.002
0.20	0.05	5 cM	1	5	0.884	0.879	0.880	0.880	0.719	0.852	0.888	10 cM	50	0.908	-0.024
0.20	0.1	5 cM	1	5	0.565	0.563	0.564	0.565	0.445	0.559	0.603	10 cM	50	0.614	-0.049
0.20	1.0	20 cM	7	15	0.021	0.017	0.019	0.019	0.019	0.020	0.021	20 cM	50	0.019	0.003
0.50	0.0	5 cM	2	10	0.997	0.997	0.997	0.997	0.884	0.971	0.990	10 cM	50	0.998	-0.000
0.50	0.05	5 cM	2	10	0.992	0.992	0.992	0.992	0.864	0.963	0.983	10 cM	50	0.994	-0.002
0.50	0.1	5 cM	2	10	0.983	0.982	0.982	0.983	0.836	0.950	0.972	10 cM	50	0.987	-0.004
0.50	1.0	10 cM	1	5	0.176	0.171	0.176	0.181	0.157	0.198	0.215	10 cM	50	0.195	-0.019

Table 4.7–continued.

Resource N=20,000, ASP=100															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	20 cM	1	10	1.000	0.999	0.999	0.999	0.998	1.000	1.000	20 cM	100	1.000	-0.000
0.01	0.05	20 cM	10	25	0.026	0.023	0.025	0.026	0.023	0.026	0.028	20 cM	100	0.023	0.002
0.01	0.1	20 cM	10	25	0.021	0.017	0.018	0.019	0.017	0.020	0.022	20 cM	100	0.019	0.002
0.01	1.0	20 cM	12	30	0.020	0.017	0.018	0.019	0.019	0.022	0.023	20 cM	100	0.018	0.002
0.20	0.0	20 cM	1	10	0.999	0.999	0.999	0.999	0.996	0.999	0.999	20 cM	100	1.000	-0.000
0.20	0.05	10 cM	2	10	0.995	0.994	0.994	0.994	0.965	0.990	0.994	20 cM	100	0.982	0.012
0.20	0.1	10 cM	2	10	0.898	0.898	0.899	0.900	0.835	0.894	0.906	20 cM	100	0.835	0.062
0.20	1.0	20 cM	12	30	0.024	0.021	0.022	0.023	0.021	0.023	0.024	20 cM	100	0.022	0.002
0.50	0.0	20 cM	1	10	1.000	1.000	1.000	1.000	0.999	1.000	1.000	20 cM	100	1.000	-0.000
0.50	0.05	20 cM	1	10	0.999	0.999	0.999	0.999	0.997	0.999	0.999	20 cM	100	1.000	-0.000
0.50	0.1	10 cM	2	10	1.000	1.000	1.000	1.000	0.993	0.999	1.000	20 cM	100	0.999	0.000
0.50	1.0	10 cM	2	10	0.396	0.380	0.385	0.390	0.353	0.416	0.439	20 cM	100	0.367	0.029

Table 4.7–continued.

Resource N=20,000, ASP=150															
Parameter		Two-stage design										One-stage design			Improv.
P.F.	k_0	map	k	n_1	F2	I.1	I.2	I.3	M.1	M.2	M.3	map	n	F1	F2-F1
0.01	0.0	20 cM	1	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	20 cM	121	1.000	0.000
0.01	0.05	20 cM	13	35	0.027	0.023	0.024	0.026	0.023	0.027	0.028	20 cM	121	0.024	0.003
0.01	0.1	20 cM	10	25	0.022	0.019	0.021	0.022	0.020	0.024	0.025	20 cM	121	0.020	0.002
0.01	1.0	20 cM	15	40	0.021	0.020	0.022	0.023	0.018	0.020	0.021	20 cM	121	0.018	0.003
0.20	0.0	20 cM	1	10	1.000	1.000	1.000	1.000	0.999	1.000	1.000	20 cM	121	1.000	0.000
0.20	0.05	10 cM	4	15	0.999	0.999	0.999	0.999	0.987	0.998	0.999	20 cM	121	0.994	0.005
0.20	0.1	10 cM	9	30	0.975	0.975	0.975	0.975	0.930	0.968	0.974	20 cM	121	0.902	0.072
0.20	1.0	20 cM	27	80	0.025	0.020	0.022	0.023	0.021	0.024	0.025	20 cM	121	0.023	0.002
0.50	0.0	20 cM	1	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	20 cM	121	1.000	0.000
0.50	0.05	20 cM	1	10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	20 cM	121	1.000	0.000
0.50	0.1	20 cM	1	10	1.000	1.000	1.000	1.000	0.999	1.000	1.000	20 cM	121	1.000	0.000
0.50	1.0	10 cM	9	30	0.564	0.559	0.562	0.565	0.512	0.574	0.592	20 cM	121	0.439	0.125

gives the that probability of the best one-stage design, i.e., with optimal m and n subject to $mn \leq N$ obtained by analytic formula. The F2−F1 column shows the increase in probability of two-stage design over one-stage.

4.4 Discussion

The Monte Carlo results indicated that the relative errors between the probabilities calculated from formulas and Markov chain simulations were almost negligible when using a 20 cM map; little discrepancy for 10 cM; a greater difference for 5 cM maps. This is most likely due to the fact that under a Markov chain model, when using a denser map, we need more samples to separate out the markers next to the markers adjacent to the genes. If we examine I.2, I.3, M.2, and M.3 columns. We will find that for a denser map it can still narrow the gene location to an area wider than the interval between two adjacent markers. For example, for N=10,000, ASP=50, P.F.=0.01, and $k_0 = 0$, M.1 is only 0.88 but M.2 is 0.97. This shows that if a denser map is used, we might need more ASPs to pinpoint the gene location.

The population frequency of the disease genes and relative risk k_0 of carrying no gene verses carrying two recessive genes at one locus have interaction on the probability of finding correct markers. For given resource, ASP, and population frequency, the probability of finding correct markers is a decreasing function of relative risk k_0. If population frequency is increased, then the speed of the decrease in probability in relative risk will be slower. This is a very intuitive result; when population frequency of disease is low and relative risk of to be affected for carrying no gene vs. carrying two recessive gene at one locus is high, there will be a lots of phenocopy cases in the sample. Hence, the probability to find correct markers will be small. On the other hand, when population frequency of disease is high, and the relative risk is low, there will only be a small number of phenocopy cases in the sample. Therefore, the probability to find correct markers will be large.

Two-stage approach once again provides big improvement in many cases, for example, when N=2,000, ASP=100, population frequency of disease gene=0.5, and the relative risk = 0.1, the probability of finding at least one of two disease gene loci is boosted from 30% to 80% by a two-stage design. The improvement over 40% were marked with $*$ in Table 4.7. Also once again, two-stage approach performs no better or worse than one-stage approach is observed, all the worse cases happened when all the ASP can be typed in one-stage designs.

CHAPTER 5
CONCLUDING REMARKS

Gene hunting is a difficult task, it usually requires a vast amount of resources. Two-stage approach is one way to reduce cost. This dissertation gives the optimal designs that can utilize resources very efficiently.

When a researcher wants to use the results in this dissertation, with information from other researchs or other consideration, he/she must postulate if the disease is caused by a single gene or multiple genes, the percentage of phenocopy, population frequency of disease gene, and relative risk. If funding and sample (ASPs) were already obtained, researcher should check the tables to find the combination of, the genetic model (dominant or recessive), funding condition (N), size of the sample (n), and heterogeneity (or population frequency and relative risk) that match his/her situation, and then use the design in the tables. If the funding and sample are not yet obtained, the researcher should consult with the tables and find the design that would fit his/her possible funding and sample conditions. For example, if a researcher believes that the disease is caused by a dominant gene, and think that 100 ASP will be available for typing and can only inquire funding enough to type 5,000 makers loci for ASPs, then using 125 markers and 20 ASPs in Stage I, choose 17 markers for Stage II study would be a design that provides a approximate 75% probability of finding the right location. On the other hand, if he/she can not collect more than 75 ASPs for the study, and funding for typing 10,000 markers is available, then use 275 markers and 15 ASPs in Stage I and choose 59 markers for Stage II would also provide a approximate 75% probability of finding the right location. If a 75% probability is not good enough, then the researcher has to find extra funding or more ASP for the

study, there is no other two-stage design can do better than the design obtained from the tables. If the combination of the parameters is out of the range of the tables given in this dissertation, then researcher should obtain the programs from this author to calculate the best design.

In the future, extending the method presented in this dissertation to searching multiple complex disease genes (more than two), a simple quantitative trait loci (QTL) and complex QTL's, and to incorporate Risch's risk ratio method are planned. Extension to incorporate different genetic model should also be included.

REFERENCES

Barnard, G.A. 1949. Statistical inference. *J. R. Statist. Soc.* B11:115-139.

Bishop, D.T., and J.A. Williamson. 1990. The power of identity-by-state methods for linkage analysis. *Am. J. Hum. Genet.* 46:254-265.

Blackwelder, C.W., and R.C. Elston. 1985. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genetic Epidemiology.* 2:85-97.

Botstein, D., R.L. White, M.H. Skolnick, and R.W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32:314.

Brown, D.L., M.B. Gorin, and D.E. Weeks. 1994. Efficient strategies for genomic searching using the affected-pedigree-member method of linkage analysis. *Am. J. Hum. Genet.* 54:544-552.

Campbell, M.A., and R.C. Elston. 1971. Relatives of probands: Models for preliminary genetic analysis. *Ann. Hum. Genet.* 35:225-237.

Carter, T.C., and D.S. Falconer. 1951. Stocks for detecting linkage in the mouse and the theory of their design. *J. Genet.* 50:307-323.

Chakravart, A., J. Badner, and C.C. Li. 1987 Test of linkage and heterogeneity in Mendelian disease using identity by descent score. *Genetic Epidemiology.* 4:255-266.

Chotai, J. 1984. On the LOD score methods in linkage analysis. *Ann. Hum. Genet.* 48:359-378.

Conneally, P.M., and M.L. Rivas. 1980. Linkage analysis in man. In: Advances in Human Genetic. New York. 209-266.

Correns, C. 1900. G. Mendels Regel über das Verhalten der Nachkommenschaft der Rassenbartarde, *Ber. dt. bot. Ges.*, 18:158-168. (Reprinted in 1950 as "G. Mendel's law concerning the behavior of progeny of varietal hybrids" in *Gentics, Princeton.* 35: suppl. pp. 33-41).

Clerget-Darpoux, F., C. Bonaiti-Pellié, and J. Hochez. 1986. Effects of misspecifying genetic parameters in LOD score analysis. *Biometrics.* 42:393-399.

Darvasi, A., and M. Soller. 1994. Optimum spacing of genetic markers for determining linkage between marker loci and quantitative trait loci. *Theor. Appl. Genet.* 89:351-357.

Davies, J.L., Y. Kawaguchi, S.T. Bennett, J.B. Copeman, et al. 1994. A genome-wide search for human type 1 diabetes susceptibility genes. *Nature.* 371:130-136.

Day, N.E., and M.J. Simons. 1976. Disease susceptibility genes–their identification by multiple case family studies. *Tissue Antigens.* 8:109-119.

Department of Energy (DOE). 1992. Human Genome: 1991-92 Program Report. Washington, D.C., US DOE, Office of Health and Environmental Research.

Elston, R. 1992. Designs for the global search of the human genome by linkage analysis. In: Proceedings of the 16th International Biometric Conference: Hamilton, New Zealand, December 7-11. Ruakura Agricultural Center, Hamilton, New Zealand, pp:39-51.

Elston, R. 1994. P-values, power and pitfalls in the linkage analysis of psychiatric disorders. In: Genetic Approaches to Mental Disorders. E.S. Gershon, C.R. Cloninger, and J.E. Barrett (eds.). Washington, D.C.: American Psychiatric Press.

Elston, R.C., and Lange, K. 1975. The prior probability of autosomal linkage. *Ann. Hum. Genet.* 38:341-350.

Faraway, J.J.. 1993. Improved sib-pair linkage test for disease susceptibility loci. *Genetic Epidemiology.* 10:225-233.

Felsenstein, J. 1979. A mathematically tractable family of genetic mapping functions with different amount of interference. *Genetics.* 91:769-775.

Gershon, E.S., L.R. Goldin, M.M. Martinez, and M.R. Hoehe. 1994. Detecting discrete genes for susceptibility to bipolar disorder or Schizophrenia. In:Genetics Approaches to Mental Disorders., E.S. Gershon, C.R. Cloninger, and J.E. Barrett (eds.). Washington, D.C.: American Psychiatric Press.

Green, J.R., and J.C. Woodrow. 1977. Sibling method for detecting HLA-linked genes in disease. *Tissue Antigens.* 9:31-35.

Griffiths, A.J.F., J.H. Miller, D.T. Suzuki, R.C. Lewontin, and W.M. Gelbart. 1993. An Introduction to Genetic Analysis. New York: W.H. Freeman and Company.

Haldane, J.B.S. 1919. The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. Genet.* 8:299-309.

Haseman, J.K. and R.C. Elston. 1972. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics.* 2:3-19.

Hodge, S.E. 1984. The information contained in multiple sibling pairs. *Genetic Epidemiology.* 1:109-122.

Holmans, P. 1993. Asymptotic properties of affected-sib-pair linkage analysis. *Am. J. Hum. Genet.* 52:362-374.

Holmans, P., and N. Craddock. 1997. Efficient strategies for genome scanning using maximum-likelihood affected-sib-pair analysis. *Am. J. Hum. Genet.* 60:657-666.

Karlin, S., and U. Liberman. 1994. Theoretical recombination processes incorporating interference effects. *Theoretical Population Biology.* 46:198-231.

Karlin, S., and U. Liberman. 1979a. Classifications and comparisons of multilocus recombination distribution. *Proc. Nat. Acad. Sci. USA.* 75:6332-6336.

Karlin, S., and U. Liberman. 1979b. A natural class of multilocus recombination processes and related measures of crossover interference. *Advance. Appl. Prob.* 11:479-501.

Klug, W.S. and M.R. Cummings. 1997. Concepts of Genetics, fifth ed. Upper Saddle River, New Jersey: Prentic-Hall, Inc.

Knapp, M., S.A. Seuchter, and M.P. Baur. 1994a. Linkage analysis in nuclear families. 1: Optimality criteria for affected sib-pair tests. *Hum. Hered.* 44:37-43.

Knapp, M., S.A. Seuchter, and M.P. Baur. 1994b. Linkage analysis in nuclear families. 2: Relationship between affected sib-pair tests and LOD score analysis. *Hum. Hered.* 44:44-51.

Kosambi, D.D. 1944. The estimation of map distances from recombination values, *Ann. Eugen.* 12:172-175.

Lander, E.S., and D. Botstein. 1986. Mapping complex genetic traits in humans: New methods using a complete RFLP linkage map. *Cold Spring Harbor Symposia on Quantitative Biology.* 51:49-62.

Lander, E.S., and D. Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics.* 121:185-199.

Lander, E.S., and N.J. Schork. September 30, 1994. Genetic dissection of complex traits. *Science.* 265:2037-2048.

Lander E.S., and L. Kruglyak. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics.* 11:241-247.

Lange, K. 1986a. The affected sib-pair method using the identity by the state relations. *Am. J. Hum. Genet.* (Letter to the Editor) 39:148-150.

Lange, K. 1986b. A test statistic for the affected-sib-set method. *Ann. Hum. Genet.* 50:283-290.

Lehmann, E.L. 1986. Testing Statistical Hypotheses. New York: John Wiley & Sons.

Levine, J.S., and K.R. Miller. 1991. Biology: Discovering Life. Lexington, Massachusetts: D. C. Heath and Company.

Lewin, B. 1990. Genes IV. Oxford: Oxford University Press.

Li, C.C., and L. Sacks. 1954. The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics.* 10:347-360.

Liberman, U., and S. Karlin. 1984. Theoretical models of genetic map functions. *Theoretical Population Biology.* 25:331-346.

Ludwig, W. 1934. Über numerische beziehungen der crossover-werte untereinander, Z. *Inder. Abst. Vererb.* 67:58-95.

Luo, D.F., M.M. Bui, A. Muir, N.K. Maclaren, G. Thomson, and J.X. She. 1995. Affected-sib-pair mapping of a novel susceptibility gene to insulin-dependent diabetes mellitus (IDDM8) on chromosome 6q25-q27. *Am. J. Hum. Genet.* 57:911-919.

MacLean, C.J., D.T. Bishop, S.L. Sherman, and S.R. Diehl. 1993. Distribution of LOD scores under uncertain mode of inheritance. *Am. J. Hum. Genet.* 52:354-361.

McPeek, M.S. 1997. An introduction to recombination and linkage analysis. In: Genetic mapping and DNA sequencing, T. Speed. and M.S. Waterman (eds.).. New York: Springer-Verlag.

Morgan, T.H. 1911. Random segregation versus coupling in Mendelian inheritance. *Science.* 34:384.

Morton, N.E. 1955. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* 7:277-318.

Morton, N.E. 1956. The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. *Am. J. Hum. Genet.* 8:80-96.

Ott, J. 1983. Linkage analysis and family classification under heterogeneity. *Ann. Hum. Genet.* 47:311-320.

Ott, J. 1991. Analysis of Human Genetic Linkage, revised ed. Baltimore: The Johns Hopkins University Press.

Ott, J. 1994. Choice of genetic models for linkage analysis of psychiatric traits. In: Genetic Approaches to Mental Disorders. E.S. Gershon, C.R. Cloninger, and J.E. Barrett (eds.). Washington, D.C.: American Psychiatric Press.

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.* 5(50):157-172.

Penrose, L.S. 1935. The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Ann. Eugenics.* 6:133-138.

Penrose, L.S. 1938. Genetic linkage in graded human characters. *Ann. Eugenics.* 8:233-237.

Penrose, L.S. 1950. Data for the study of linkage in man: Red hair and the ABO locus. *Ann. Eugenics.* 15:243-247.

Penrose, L.S. 1953. The general purpose sib-pair linkage test. *Ann. Eugenics.* 18:120-124.

Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. 1992. Numerical Recipes in C, second edition. New York: Cambridge University Press.

Rhodes, F.H.T., R. Merrilees, and R. Zallinger. 1974. Evolution. New York: Golden Press.

Rao, C.R. 1973. Linear Statistical Inference and its Applications, 2nd ed. New York: John Wiley & Sons.

Rao, D.C., Morton, N.E., Lindsten, J., Hulten, M., and Yee, S. 1977. A mapping function for man. *Hum. Hered.* 27:99-104.

Risch, N. 1990a. Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.* 46:222-228.

Risch, N. 1990b. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. *Am. J. Hum. Genet.* 46:229-241.

Risch N. 1990c. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected related pairs. *Am. J. Hum. Genet.* 46:242-253.

Risch, N. 1992. Corrections to "Linkage Strategies for Genetically Complex Traits. III. The Effect of Marker Polymorphism on Analysis of Affected Relative Pairs" (AM. J. Hum. Genet. 46:242-253, 1990). *Am. J. Hum. Genet.* (Letter to the Editors) 51:673-675.

Risch, N., and K. Merikangas. September 13, 1996. The future of genetic studies of complex human diseases. *Science* 273:1516-1517.

Schaid, D.J., and T.G. Nick. 1990. Sib-pair linkage tests for disease susceptibility loci: Common tests vs. the asymptotically most powerful test. *Genetic Epidemiology.* 7:359-370.

Self, S.G., and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Stat. Assoc.* 82(398):605-610.

Serfling, R.J. 1980. Approximation Theorems of Mathematical Statistics. New York: John Wiley & Sons.

Shah, S., and J.R. Green. 1994. Disease susceptibility genes and the sib-pair method: A review of recent metodology. *Ann. Hum. Genet.* 58:381-395.

Shute, N.C.E. 1988. Power properites of the likelihood ratio test in an analysis of genetic diseases. *Biometrics.* 44:951-958.

Smith, C.A.B. 1963. Testing for heterogeneity of recombination values in human genetics. *Ann. Hum. Genet.* 27:175-182.

Speed, T.P. 1997. What is a genetic map function? In: Genetic Mapping and DNA Sequencing, T. Speed and M.S. Waterman (eds.). New York: Springer-Verlag.

Sturt, E. 1976. A mapping function for human chromosomes, *Ann. Hum. Genet.* 40:147-163.

Suarez. B.K., J. Rice, and T. Reich. 1978. The generalized sib pair IBD distribution: Its use in the detection of linkage. *Ann. Hum. Genet.* 42:87-94.

Tan, W.Y., and W.C. Chang. 1972. Convolution approach to the genetic analysis of quantitative characters of self-fertilized populations. *Biometrics.* 28:1073-1090.

Tan, W.Y., and H. D'Angelo. 1979. Statistical analysis of joint effects of major genes and polygenes in quantitative genetics. *Biometrical J.* 21(2):179-192.

Weeks, D.E., and K. Lange. 1988. The affected-pedigree-member method of linkage analysis. *Am. J. Hum. Genet.* 42:315-326.

Weeks, D.E., and K. Lange. 1992. A multilocus extension of the affected-pedigree-member method of linkage analysis. *Am. J. Hum. Genet.* 50:859-868. b

Weir, B.S. 1996. Genetic Data Analysis II. Sunderland, Massachusetts: Sinauer Associates, Inc.

BIOGRAPHICAL SKETCH

Chi-Hse Teng was born and grew up in Taipei, Taiwan, Republic of China. He received his B.S. in mathematics from National Taiwan University in 1987. He servered as a military policeman in the Chinese Army from 1987 to 1989. After an honorable discharge, was a research assistant at the Institute of Statistics, Academia Sinica, R.O.C., before beginning his graduate study in the Department of Statistics at the University of Florida. He won a prize in the student submission competition of the Third Annual Florida Epidemiology Meeting in July 1997.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Mark C.-K. Yang, Chairman
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Randolph L. Carter
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Frank G. Martin
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Susan P. McGorrary
Research Assistant Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jin Xiong She
Associate Professor of Pathology and
Laboratory Medicine

This dissertation was submitted to the Graduate Faculty of the Department of Statistics in the College of Liberal Arts and Sciences and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

December, 1997

Dean, Graduate School

LD
1780
1997
.T292